DOCUMENT RESUME

ED 093 989 TM 003 845

AUTHOR	Klitgaard, Robert E.
TITLE	Achievement Scores and Educational Objectives.
INSTITUTION	Rand Corp., Santa Monica, Calif.
SPONS AGENCY	National Inst. of Education (DHEW), Washington, D.C.
REPORT NO	R-1217-NIE
PUB DATE	Jan 74
NOTE	73p.
EDRS PRICE	MF-$0.75 HC-$3.15 PLUS POSTAGE
DESCRIPTORS	Academic Achievement; *Achievement Tests; *Educational Improvement; *Educational Objectives; Evaluation Needs; *Evaluation Techniques; Scores; Social Background; *Test Results

ABSTRACT

This report deals with new ways to look at achievement test scores as measures of educational outcomes. It attempts to show how existing data might be used more productively by planners and evaluators, by suggesting how test score statistics other than the mean might be used, and how they might indicate success along a wide variety of educational objectives. The report is a theoretical effort in which the empirical behavior of the proposed new statistics is examined briefly. (BB)

ED 093989

TM 003 845

ACHIEVEMENT SCORES AND EDUCATIONAL OBJECTIVES

PREPARED FOR THE NATIONAL INSTITUTE OF EDUCATION

ROBERT E. KLITGAARD

R-1217-NIE
JANUARY 1974

U.S. DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Rand
SANTA MONICA, CA. 90406

This research was supported by the National Institute of Education under Contract No. B2C-5326. Reports of The Rand Corporation do not necessarily reflect the opinions or policies of the sponsors of Rand research.

Published by The Rand Corporation

PREFACE

How should achievement test scores be used to evaluate public schools and educational programs? This study explores this question and suggests some ways in which statistics other than test score averages can be useful to educational planners and evaluators. The study was conducted in support of Rand's Analysis of the Education Voucher demonstration, which is currently taking place in Alum Rock, California. The demonstration is a large-scale social intervention with a wide range of objectives, including increased parental influence and satisfaction with schools, greater responsiveness by educators to the needs of parents and children, more diversity of educational programs, and, ultimately, better education for children. Funded initially in 1972 by the Office of Economic Opportunity, it is currently supported by the National Institute of Education. Rand has been studying the demonstration since its inception in order to assess its effects on relevant aspects of the social, political, economic and educational systems in the Alum Rock School District. In addition to a wide variety of analyses related to the demonstration itself, Rand has undertaken a number of studies of special topics, designed to enhance our ability to provide national policy advice in this area. This report presents the findings of one such special study.

Many educators believe that the way in which achievement data are now typically used for evaluation is unsatisfactory. One criticism contends that tests lack high reliability and validity. Rand researchers are studying new methods of testing that promise some improvements in these areas.* Other criticisms challenge the relevance for education policy decisions of even perfectly valid and reliable test scores. Achievement tests, it is argued, measure only

---

*Emir H. Shuford and Thomas A. Brown, *Elicitation of Personal Probabilities and Their Assessment*, R-1371-ARPA (forthcoming); and W. L. Sibley, *An Experimental Implementation of Computer-assisted Admissible Probability Measurement*, R-1258-ARPA (forthcoming).

some of the multiple objectives of schooling, and evaluations using only achievement measures are misleading, and perhaps even irrelevant to the true goals of education. Unfortunately, achievement tests are almost the sole source of data available, and policy decisions often cannot wait until new, better measures are created and instituted.

This report attempts to develop new ways to look at achievement test scores as measures of educational outcomes. It tries to show how existing data might be used more productively by planners and evaluators, by suggesting how test score statistics other than the mean might be computed, and how they might indicate success along a wide variety of educational objectives. The report is primarily a theoretical effort in which the empirical behavior of the proposed new statistics is examined only briefly. Future work, based in part on data collected during Rand's study of the voucher demonstration, will explore many of these issues in more detail.

SUMMARY

In theory, an evaluation requires precisely defined objectives, valid and reliable measures, and some mathematical expression ("objective function") relating the measures to the objectives. But in public schools, objectives are fuzzy and controversial; measures are nonexistent, except for achievement tests; and evaluations implicitly assume that schools with higher average achievement scores are better--implying a simplistic and probably incorrect objective function. For evaluators, education is the worst of worlds, but evaluations must nonetheless be made.

Even though they relate only to some school objectives, achievement scores should be used, because some information is better than none. But evaluations should go beyond *average* scores and look at statistics of intraschool *distributions* of scores, because more information is better than some.

What additional statistics should be examined? This question asks for a specification of the objective function--a methodologically tractable task, but one that would be infeasible given the realities of the educational system. One can, however, provide a number of easily computable statistics of the intraschool distribution of both *uncontrolled* and *residual* achievement scores that have intuitive links to ill-defined but still meaningful educational objectives. For example:

Objective	Measure
General achievement level	Mean
Achievement relative to student background	Residual mean
Equality of achievement	Spread
Equalizing effect of school	Actual minus expected spread
Mobility afforded by school	Residual spread
Effectiveness with exceptional children	Distortion

Objective	Measure
Effectiveness with over- and underachievers	Residual distortion
Assuring children achievement skills at minimum level K	Proportion of students' scores above K
Assuring children do not underachieve below level C	Proportion of students' residuals above C
Success with children above (below) background level S	Mean score of children above (below) S

These crude measures should be used crudely, not cardinally. They are probably not preferentially independent; but efforts at concocting some one grand measure of school success would be ill-advised. Precisely which estimator should be chosen for each measure (for example, the variance, interquartile range, or relative mean deviation for spread) is a matter for further investigation, but what really matters is that some such estimators be considered when evaluating.

Preliminary examinations of some of the new measures reveals that nonschool background factors do not explain interschool variation along those measures as well as these factors explain variations in school mean scores. Also, some schools seem to equalize student scores far more than chance alone would predict.

Introducing the new measures would be a step away from simplistic evaluations and adverse incentives for educators--and perhaps a step toward organizational change.

ACKNOWLEDGMENTS

Many of the ideas explored in this report stemmed from conversations with George Hall. Franklin Berger performed most of the computer applications. Theodore Donaldson, Gus Haggstrom, and Richard Zeckhauser commented on earlier drafts, leading to major improvements (although perhaps not as many as they hoped). I also profited from discussions with Emmett Keeler, Milbrey McLaughlin, and Michael Spence. The usual caveat protecting these courteous people from further responsibility is, of course, in order.

CONTENTS

PREFACE iii

SUMMARY v

ACKNOWLEDGEMENTS vii

ANALYTICAL PRECIS xi

Section
I. INTRODUCTION 1

II. OBJECTIVES MISSED BY MEAN SCORES 5

III. SPECIFYING OBJECTIVE FUNCTIONS: THE THEORY VERSUS EDUCATIONAL REALITIES 12
Objective Functions for Evaluation 12
From Theory to Practice in Education 16

IV. MEASURES FOR EDUCATIONAL EVALUATION 20
Measures of Central Tendency 20
Spreads 22
Distortions 41
Proportions Above Certain Thresholds 44

V. PRACTICAL CONSIDERATIONS AND CONCLUSIONS 47

REFERENCES 51

ANALYTICAL PRECIS

Roman numerals refer to sections; arabic numbers correspond to those in the margins of the paper.

I. Introduction

1. Evaluations in education have come under fire
2. for a number of reasons (largely, perhaps, because the evaluations bear unpleasant results). Especially vulnerable has been the use of cognitive achievement scores to measure "success."
3. Achievement scores have many imperfections, especially their lack of congruence with educational objectives,
4. but it is not clear what can take their place.
5. Given that achievement scores are the only currently available measure, how can such scores best be used in evaluation?
6. This report outlines some improvements in the use of achievement data, particularly the implications of moving beyond the school or program *mean* to examine the *distribution* of scores.

II. Objectives Missed by Mean Scores

7. Choosing the right measure for evaluation depends on program objectives.
8. Even if one has the right measure, one may use the wrong (or insufficient) statistics.
9. For example, consider the evaluation of a country's economic well-being. Suppose income is the right measure. Besides the average (or per capita) income, statistics of income distribution are important; so are statistics indicating the well-being of certain sub-groups and the number of citizens below a defined poverty line.
10. In investment decisions, too, the mean is not a sufficient evaluator, even if profit is the correct measure.
11. Educational evaluation should also go beyond the mean.
12. Unfortunately, in education, achievement scores are the only widely available data,
13. but one can use both uncontrolled and residual measures (where nonschool factors are held constant).
14. Whether one chooses uncontrolled or residual scores (or both) as the *measure*, the use of the school or program mean as the *statistic* of interest implies a number of probably untenable

assumptions about educational objectives. Theoretically, to decide which statistics are best for evaluation, one needs a well-specified objective function.

III. Specifying Objective Functions: The Theory versus Educational Realities

15. Specifying an objective function for education involves three questions:

16. (1) How does one compare students' scores of 35, 40, and 45? Does one value all five-point gains the same? Many plausible objective functions do not.

17. (2) How can one combine an evaluation of a student's uncontrolled score with an evaluation of his achievement relative to his nonschool background?

18. (3) How can evaluations for individual students be combined to obtain a school or program index of success?

19. Using the school mean as the only evaluation measure implies unreasonable answers to these questions.

20. Theoretically, a rational decisionmaker could produce the answers and, thereby, the appropriate "statistics" for evaluation.

21. But given educational realities, the methodology of utility functions is probably inapplicable

22. because local educational jurisdictions and decisionmakers have differing objectives, and

23. obtaining utility functions from the relevant parties is infeasible.

24. Statistics for evaluation, then, cannot be deduced from a given educational objective function,

25. but the present reliance on the mean as the statistic is inadequate.

26. What should be done? Recall the analogy from income distribution: not everyone agrees on the best statistics, but this fact doesn't stop *some* indicators of distribution and poverty from being used--and being useful.

IV. Statistics for Educational Evaluation

27. The *central tendency* of a school or program's performance is not everything one wants to know, but it is important for evaluation.

28. Even to measure the central tendency, however, the mean may not be the best statistic. The choice of statistics depends on the degree of "robustness" desired

29. and the choice of fitting techniques,

30. as well as on considerations of pure convenience.

31. Whether the residual mean is derived from individual or already aggregated data also matters.

32. Fortunately, these imperfections are not critical, since one needs only a proxy for a loosely defined educational goal. Precision here is an illusion.

33. School mean scores and residual mean scores are distributed relatively normally across schools. Residual mean scores do not correlate highly over time; but the residual mean correlates rather highly within random halves of the same school in the same year.

34. Equality of educational outcomes is an increasingly voiced goal; the *spread* of a school's scores is a good proxy for this objective.

35. Since the spread of a school's scores depends in part on nonschool factors, a second measure will be needed to assess the school's equalizing ability: the difference between its observed and expected spread.

36. A third measure of spread, based on student-level regressions, seems a promising indicator for the amount of mobility (both upward and downward) that a school provides.

37. Which particular statistic of spread should be used for each of the three measures? The answer awaits further empirical work; and probably, as in the case of income distribution, no one statistic will be universally accepted as optimal.

38. Empirical explorations showed that school standard deviations and the difference between observed and expected standard deviations have negatively skewed distributions across schools. Nonschool background factors could not explain the variation among school standard deviations nearly as well as they can the variation among school means. Some schools consistently equalized outcomes, even after controlling for socioeconomic status, racial composition, size of school, and school mean score—in fact, averaging over time a reduction of 20-25 percent in variability of outcomes compared with the typical school.

39. Insofar as educational policy emphasizes slow or fast learners, neither the mean nor the spread of a school's score will be an efficient evaluative statistic.

40. The *distortion* of a school's scores, measured perhaps by the skewness of the school's distribution of results, may be a useful statistic.

41. So might the distortion of a school's distribution of residuals from student-level regressions.

42. The skewness statistic presents analytical difficulties but has a tradition of use in economics in analogous circumstances.

43. Statistical problems also attend the skewness statistic's use,

44. but, other things equal, the more positively skewed a school's distribution of scores (residuals), the better it is doing with slow and fast learners (over- and under-achievers), although at the expense of its average students.

45. If some particular level of attainment is of interest, the *proportion* of a school's students above that level is an appropriate evaluative statistic.

46. Such threshold definitions of success are often implied in education.

47. Thresholds of both uncontrolled and residual scores are useful.

48. In one empirical exploration, nonschool factors explained school mean scores well but not the scores of students with IQs above 123 or below 93.

V. Practical Considerations and Conclusions

49. There are three undesirable alternatives in educational evaluation: to ignore achievement scores altogether, to rely on school or program means alone, or to insist on perfect statistics that are possible only given unobtainable objective functions.

50. This study suggests a second-best course of action: to use easily computable measures for a large number of ill-defined but still important educational objectives.

51. Which exact statistics to use is a question for further research; but some such statistics are better than none.

52. In practice, the new statistics should be used crudely, dividing schools into five groups along each statistic. One should resist the temptation to combine the statistics into an overall indicator of school success.

53. Bureaucratically, the use of additional statistics may avoid adverse incentives and stimulate welcome change.

54. Further research is in order; but even in their present form, the new statistics can improve educational evaluation.

I. INTRODUCTION

1. Evaluations are themselves subject to evaluation, especially when they bear bad news. In public education, where most evaluations have failed to show consistent and important relationships between what goes on in schools and variations in student learning (Averch et al., 1972; Jencks et al., 1972), the evaluations have themselves received bad marks from numerous critics. "We need only to look at the large mass of 'no significant difference' fundings typically produced by evaluation studies to begin to wonder about the power of the techniques [for evaluation], particularly when all the evidence of the senses of the participants argues that there is a difference" (Guba, 1967, pp. 58-59). A recent review of ESEA Title I despairs: "If the evaluations presently being done are a yardstick of what has been learned from seven years and over 50 million dollars of Title I evaluation, the conclusion must be that we have learned very little" (McLaughlin, 1973, p. 180).

2. Many reasons have been cited for the failure of evaluation in education--as opposed to the failure of education--including the politicization of research, inefficient design, too short time frames, high student turnover, insufficient funds, inadequate federal appetite for results of evaluations, misconceived analogies from economics, adverse local incentives, too microscopic levels of analysis, and no doubt others. But no criticism is more common than attacks on the use of achievement test scores to measure educational success.

Achievement tests have few friends these days. Long the darling of educational researchers, axiomaticians, and "human engineering" enthusiasts, standardized testing is now widely scorned, scolded, and even sued--not least by educators themselves. It has been a striking change of tune.

During the 1950s and 1960s, there was an explosion of specialized research in the construction of more and better tests, and many social activists felt that only with the large-scale use of objective measures like tests could the inequalities in the educational system be publicly recognized and confronted. Testing happily combined mathematical

muscle and redistributional relevance--the quintessentially desirable academic mix. But when the numbers begin to imply that sociological moralisms might be misplaced, even that "education" might not be educational, criticisms of the quantification proliferated. The objections have been couched in mathematical and methodological terms, but they stem more from the unexpected and unpleasant implications of studies using achievement tests than from any newly discovered statistical shortcomings of the tests themselves.

3. The criticisms of achievement tests fall into three categories: the tests' validity and reliability, imperfections in data collection and aggregation, and the relationship of test scores to educational goals. On the first point, the testers themselves, perhaps embarrassed by the sudden seriousness with which their progeny were being taken by policymakers, called IQ and grade equivalency scores "monstrosities."[1] Second, large governmental data collections and surveys--compilations so long begged for by educational researchers and, indeed, designed by their peers and leaders--were now decried as incomplete, biased, plagued by attrition and turnover: in short, as no basis whatever for policy. (For example, Bowles and Levin, 1968; Hanushek and Kain, 1972; and Guthrie, 1972.) And third, many critics said that achievement scores were simply the wrong measure of output: they encompassed only a small part of what schools were about, and that part imperfectly.

The first two objections are doubtlessly important, but they will not be dealt with here. The concern of this report is the third type of criticism, pointing to the lack of congruence between achievement measures and educational objectives. Critics remind us that the goals of education are much broader than just reading and mathematics and call for a renewed appreciation of the human and social values of schooling. It is often implied that studies relying exclusively on achievement tests should be relegated to the status of interesting but irrelevant academic artifacts.

---

[1] Henry S. Dyer, Vice President of the Educational Testing Service, cited in Stevens (1971).

4. Unfortunately, it is not clear what can take the place of achievement data for evaluation purposes. Not only are the objectives of public education unclear, the measures that might be used to gauge progress in such areas as "affective growth," "self-concept," "ability to cope," and others are little developed and of doubtful operational usefulness (Wargo et al., 1971). Other measures of school performance--drop-out rates, proportion of students entering college, absenteeism and vandalism rates, and so forth--are often useful, but unfortunately they are seldom uniformly available over time and space. The criticisms of achievement scores have often stopped short of recommending what course evaluation policy should take in education--apart, of course, from the inevitable calls for more research on objectives and better measures.

5. If evaluative decisions are nonetheless necessary, perhaps one should adopt a suboptimizing approach. *Given* the lamentable fact that achievement scores are the only widely available measures, how can they best be used in evaluation?

6. This report outlines some possible improvements in the ways achievement data are now being used by government data banks and in large-scale evaluations. Useful information is being thrown away. When schools are examined, most evaluations look only at the *average* achievement score in each school; ditto for programs, districts, and so on. When large-scale regression analyses (like the Coleman report) are performed, they look at the *average* effect of policies across all schools. In the process, evaluations capture only the central tendency of what is happening in a school or across a population of schools. They overlook what happens to certain kinds of students and certain kinds of schools; they miss the interesting phenomena occurring on the tails of the distribution of scores; they omit the spread and shape of scores within a school or across schools. And by concentrating on averages, certain things are implied about educational objectives and the likely behavior of schools that upon a moment's reflection are almost certainly false.

This study considers the implications of moving beyond the school mean in evaluating how well a school is doing. It proposes other statistics of the intraschool distribution of scores that link test results with certain ill-defined but still meaningful goals such as equality,

educational opportunity, success with retarded and gifted children, success with children of certain social backgrounds, and minimum levels of reading and mathematical skills.

II. OBJECTIVES MISSED BY MEAN SCORES

7. How should achievement measures be used to assess success in schools?[1] The choice of measures for evaluation depends, in the first instance, on educational objectives. "The basic question is, 'what is to be measured?' This question can be answered only if a more fundamental question is asked and answered, 'For what purpose?'" (Bauer, 1966, p. 39). A standard work on evaluation reaffirms this obvious but often forgotten truth, even if in forgettable prose:

> The most identifying feature of evaluative research is the presence of some goal or objective whose measure of attainment constitutes the main focus of the research problem. Evaluation cannot exist in a vacuum. One must always ask evaluation "of what." Every action, every program has some value for some purpose--therefore, it is meaningless to ask whether a program has any value without specifying for what. (Suchman, 1967, pp. 37-38.)

Yogi Berra is more memorable, though on a slightly different subject: "If you don't know where you're going, you probably won't get there."

8. Objectives should determine not only the choice of measures but also the choice of statistics. One may have the right measure but use the wrong (or incomplete) statistics. Suppose we have data--the right data--on a school's students. How do we evaluate a school's success? The standard answer in most large-scale evaluations, whether done by governments or by private scholars, whether found in official data banks or in sporadic surveys, has been to look at the school's *average* score.[2] Even if achievement tests are the right measure, the school mean may be the wrong--or at least only a partial--statistic.

9. Compare the case of income distribution. Suppose a person's economic assets form a satisfactory measure of his welfare, either because

---

[1] The unit of analysis might not be schools, but districts, programs, counties, states, etc. For simplicity, I shall assume that the relevant unit is the school.

[2] Perhaps suitably controlled for nonschool background factors.

there are no other objectives than economic ones, or because a uniform metric of willingness to pay can translate other types of objectives into an economic measure (under stringent conditions that can be considered met), or because we are concerned for the moment with his economic welfare and our appreciation of that is independent of other dimensions of welfare. Suppose income is the metric for individuals, and the desire is to evaluate the welfare of a group--say, a country. What statistics are appropriate? Most people would maintain that the national average income would not be the only statistic of interest. To be sure, per capita income is widely used to rank nations' economic development and to indicate secular trends. But no description of a nation's economic welfare would be complete without some measure of the *distribution* of income--its dispersion among rich and poor--whether the statistic used is the variance, the coefficient of variation, Pareto's α, the Gini index, Høivik's relative mean deviation, or Lydall's percentiles.[1]

Other sorts of income dispersion might be of importance in evaluating a nation's economic progress. The relative wealth of particular groups--racial minorities, sexes, ages, and so on--would not be captured by measures of inequality for the whole society. Yet these groups might be the targets of many national economic programs, the success of which

---

[1] *Variance*: the mean of the squared differences between the individual incomes and the national average.

Coefficient of variation: the square root of the variance divided by the national average.

Pareto's α: fit the income distribution above the mode with $N = AY^{-\alpha}$, where Y = level of income, N = proportion of citizens with incomes $\geq$ Y, and A is a constant; then α is Pareto's measure of distributional equality (larger αs are more equal).

Gini index: plot the cumulative proportion of the population against the cumulative proportion of income, obtaining the Gini (or Lorenz) curve; then the Gini index is twice the area between the Gini curve and a 45° line (lower Gini indexes are more equal).

Høivik's relative mean deviation: twice the maximum distance from the Gini curve to the 45° line (lower relative mean deviations are more equal) (Høivik, 1971).

Lydall's percentiles: compare percentiles (from the bottom) 99, 98, 95, 90, 80, 25, 15, and 5 with the median income (Lydall, 1968).

Many other measures of income inequality and distribution have been suggested.

could not be gauged using the national average or some index of national income distribution.

Many assessments of economic well-being also concern themselves with poverty, usually defined using an (arbitrary) threshold below which a citizen is called poor. Generally, the mean and the dispersion alone do not reflect this concern: The statistic of interest is the proportion of the population that falls below the poverty line.[1] Economic policies that combat poverty would be poorly evaluated using only per capita income figures or changes in the Gini index.

10. For another example, consider investment decisions. Once again, suppose income received was considered the appropriate measure of successful investment. The expected value of the uncertain future income stream would probably not be the sole statistic of interest. If one is risk averse (or risk prone), some measure of the spread of the probability distribution of benefits is important. Given two distributions of payoffs with equal means and variances, if one with a lower mode but a reduced probability of large losses is preferred, then the skewness statistic should probably also be considered. A large literature in theoretical economics has grown around the inclusion of risk measures in investment decisions; its message is that the mean is not enough.[2]

11. Educational evaluations should be similarly informed about aspects of school success beyond the average score. School policies are also concerned about equality of outcomes, success with fast and slow learners, students from underprivileged backgrounds, mobility and educational opportunity, and certain minimum levels of attainment. Judging schools only on the basis of average scores overlooks all these objectives.

After we agree to go beyond just the mean score, two questions arise: (1) Beyond the average score along what *measures*? (2) Beyond to what *statistics*?

---

[1] The line may be defined absolutely (below $N for an urban family of four) or relatively (the lower M%, where the statistic of interest is obviously not the proportion but the composition and stability of the group thus defined; or below $ku, where u is the national average income and k is some proportion).

[2] See, for example, Chipman (1973) and Fisher and Hall (1969), and the references therein.

12. What measures? In education one hardly has a choice. Goals are vaguely specified in most educational systems, making the selection of worthwhile measures nearly impossible; and only a few measures even indirectly related to school objectives have been developed with precision and widely implemented. The results of cognitive achievement tests are the only data widely available, making one's choice, for the present, quite constrained.

13. Deciding which form of achievement score to use is not easy. In educational evaluation one is not trying to assess the well-being of a group, or at least not just that; one wants to evaluate the contribution of policy-related variables of the *educational* system to that well-being. For system evaluation, one might prefer a value-added or *residual* measure of achievement, not the achievement scores themselves. The reason is straightforward: Differences among the scores of students from different schools cannot be attributed entirely to the schools alone. Pupils bring different amounts of intellectual capital to their learning experiences because of differing socioeconomic, psychological, and genetic backgrounds. Schools with superior students will tend to attain superior results, but not necessarily because of superior schooling.

Therefore, many writers have called for the use of residual achievement scores to evaluate public education (for example, Barro, 1970; Dyer, 1972). Only by taking the students' varying nonschool background factors into account, they argue, can the differences between school scores be linked to the quality of the education provided. Since achievement scores apparently reflect students' background factors much more than they do school-related policy variables,[1] working with unadjusted scores will mask the effects of alternative school policies. Therefore, residual scores are required.

Residual scores have their opponents. There are a host of statistical problems, not least of which is choosing the appropriate control

---

[1] Smith, using data from the Equality of Educational Opportunity Survey, found that only between 5.85 percent and 7.46 percent of the variation among unadjusted school mean achievement scores is *potentially* due to school effects. Cited in Jencks *et al.* (1972), p. 178.

variables. At best socioeconomic measures are proxies for the background factors one wishes to hold constant across schools, and the predictive power of various controls may differ from community to community, making residual scores difficult to interpret.[1] Some argue that residual scores computed from school-level data are subject to computational unreliability.[2] Even working with individual residual scores is subject to statistical errors of many kinds.[3] If there is multicollinearity between school variables and nonschool background factors, further uncertainty is introduced into the estimation of school effects.[4]

A non-statistical, normative problem also attends the use of residual scores. Evaluating with residual scores implies that the regression line (relating background factors to achievement) is accepted as the normative baseline from which to judge policy. To some educators, the fact that the regression line indicates differences in achievement across economic classes, geographical areas, and racial groups is part of the problem and is itself an indicator of poor performance by the

---

[1] These problems are often recognized by advocates, but usually left unresolved; see, for example, Barro (1970), pp. 203-205. Dyer (1972), p. 526, concludes cheerfully:

> Anyone who examines closely the method I am proposing for assessing the educational opportunities provided by schools will find plenty of problems in it, some theoretical or technical and some practical. There is no space here to discuss these problems, but I am convinced that, possibly with some modifications of the basic model, they can be solved.

For a less sanguine view, see Cronbach and Furby (1970).

[2] Dyer, Linn, and Patton (1969), implicitly assuming that separate regressions used to control individual scores and school scores for background factors were free from error, found that school-level residuals had undesirably low correlations with aggregated individual-level residuals for the same schools.

[3] Residual variation could arise from other causes than differences in school effectiveness: imperfection in measurement, misspecification of background factors, omitted variables, poor choice of fitting technique, incomplete data, regression toward the mean, and the combined random fluctuations involved in all the regressor variables.

[4] Given multicollinearity, the significance of each affected variable will be difficult to interpret. Also, if the amount of multicollinearity varies from regression to regression, not only will significance tests be difficult, but techniques for partitioning shared variance will give different answers. See Mayeske *et al.* (1969) and Craemer (1971).

educational system. Some educators have maintained that using residual scores endorses existing inequalities as the proper frame of reference for evaluation.

The choice of measures may depend on the choice of problems one wishes to analyze. To evaluate *cost-benefit* aspects of education--to compare the educational dollar's productivity with a dollar for defense, housing, or tax refunds--one may prefer an absolute achievement measure. However, for *cost-effectiveness* questions--to compare one school or educational practice with another--a residual measure may be better. Part of the normative issue stems from the desire to do two different kinds of evaluation.

A policymaker may choose which sort of measure seems best, but there may be no need to be exclusive. Both measures are useful, and both convey different kinds of information about a school's performance. The wisest strategy, then, might be to use both unadjusted achievement data and achievement residuals. Sections IV and V discuss aspects of both the theory and the practice of managing their merger for evaluative purposes.

14. The mean is a useful summary statistic of a school's performance under certain circumstances. But using only the mean for evaluation both throws away information and makes assumptions that are probably untenable. Using the mean for evaluation implies:

(a) An increase in an achievement score of a given magnitude is valued equivalently, no matter where on the achievement scale it occurs. (A gain from 25 to 30 is just the same as a gain from 65 to 70, for example.) But the assumption is false if we care particularly about the attainment of certain basic skills, or if high scores are very desirable. Where educational policy does not equally value equal-sized gains on a standardized achievement test, the mean will not accurately reflect educational objectives.

(b) All students are valued equally (since the arithmetic mean adds all students' scores in an unweighted fashion, dividing by the total number of students). But educational policy

may attach greater weight to academic gains among certain students, perhaps to overcome past disadvantages or to increase the proportion in certain academic specialties. Insofar as a policy is directed at certain types of students, the mean school score will not be adequate for evaluation.

(c) Student i's score is independent of student j's (the mean merely sums scores, without adjusting individual scores depending on the scores of others). This assumption may be false for two reasons. First, one may care about the *distribution* of scores across students: the equality of outcomes, the amount of mobility, the riskiness of educational outcomes, the tails of the distributions of scores. The mean does not communicate the distribution, just its central tendency; the analogue to income distribution is obvious. Second, if education acts as a screening device or filter for later education or for the job market, scores i and j cannot be treated as if they were independent.

One suspects, therefore, that evaluations using only mean school scores may be missing some important objectives of education. To decide which additional or substitute statistics would be appropriate, one must return to the question of goals in a more rigorous fashion.

III. SPECIFYING OBJECTIVE FUNCTIONS: THE THEORY VERSUS EDUCATIONAL REALITIES

OBJECTIVE FUNCTIONS FOR EVALUATION

15. An objective function is the formal link between objectives and evaluative measures. The idea behind an objective function is to assign a numerical value (utility) to every (relevant) state of the world; the decision problem is to maximize that function subject to budget and operational constraints. With such a function a school or program can be evaluated merely by examining its utility score and the costs of attaining that score.

To construct an objective function for achievement scores, three questions require answers:

(a) How does one evaluate one achievement score compared with another (or one residual score compared with another)? We may tautologically define some objective function $U_A = f(A)$, where A signifies the achievement score, or some function $U_R = g(R)$, where R signifies the residual score, but what do the functions f and g actually look like?

(b) How could U_A and U_R be combined into a single, composite objective function U_T for each student?

(c) If one is evaluating schools and not students, how could the U_{Ti} be combined for each student i into a school index?

16. Question (a). How does one compare scores of 35, 40, and 45? We know that 35 is five points lower than 40, and 40 five points lower than 45. But the units here are derived through some standardization process used by the testers, norming scores to some population of students. There is no necessary reason why this scale should correspond to one's evaluation of those scores. Does one equally value a five-point increase whether it is from 35 to 40 or from 40 to 45 (or from 60 to 65)? To answer this question a utility function for an individual's score is required.

Theoretically, the evaluator could construct this utility function by presenting the decisionmaker with choices between lotteries on scores. For instance, is it better for a student to have a score of 50 for sure or a 50-50 lottery on scores of 40 and 75? If you were indifferent, your utility function for the student's achievement could be suspected of being convex over that region. In the well-known von Neumann-Morgenstern fashion, a set of lottery questions could ascertain the entire utility function of a rational decisionmaker.[1]

It is difficult to predict what utility function for achievement scores would be specified. Decisionmakers might well disagree. One answer--though in my opinion unlikely--is that in fact a five-point achievement score increase would be weighted the same whether it were from 35 to 40 or 60 to 65 or anywhere else. In such a case, U_A would be some linear function of the score, as in Fig. 1a.

Another observer might consider increases in low scores more valuable than gains in scores that are already high. If questioned in detail about his preferences for a student's scores, this observer might respond with a U_A curve like the one in Fig. 1b.

If one valued achievement gains on both the low and high ends more than those in the middle--perhaps because of an emphasis on slow learners and the gifted--a cubic utility function like Fig. 1c might be the appropriate representation.

Suppose one's educational objective were predominantly to ensure that the student achieved a score above some minimum level k--perhaps some threshold of needed cognitive skills. Achievement increases beyond k are relatively unimportant. Then a modified step-function like Fig. 1d would be the right utility function to use for evaluation.

Clearly the shape of U_A might be many things besides linear. Different policymakers might choose different functions; different programs might want to weight achievement gains differently; and utility functions might vary for different kinds of students. Similar remarks

---

[1] von Neumann and Morgenstern (1944). See also Friedman and Savage (1948); a lucid elementary exposition is found in Raiffa (1968), Ch. 4. Roche (1971) had local educational administrators make explicit their utility functions for different kinds and levels of student achievement scores.

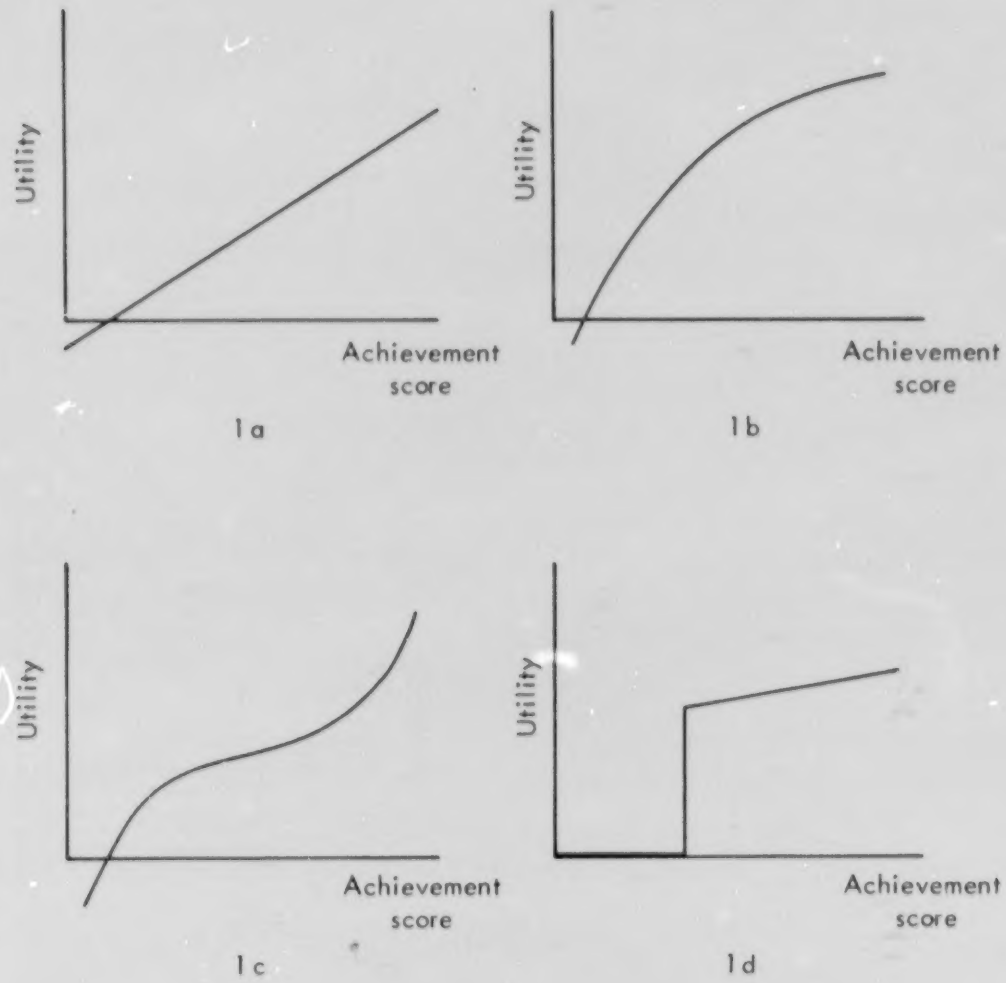


Fig. 1 — Some plausible shapes for utility functions for achievement

apply for U_R: *a priori* it seems unlikely that g(R) should be linear, and no other shape recommends itself as the obvious alternative.

17. Question (b). Suppose we have elicited U_A and U_R. How can we combine them into some overall utility function U_T? Theoretically, to answer this question one first assesses the interdependence of the two functions. Does our evaluation of U_A for student i depend on his residual score? That is, is the choice among lotteries on achievement scores any function of the student's residual score, or vice versa? If we hold the residual score fixed at some level R_0, do our conditional (probabilistic) preferences for the unadjusted score A depend on what fixed value R_0 is chosen, and vice versa? If *not*, then the composite utility function U_T has an additive representation:[1]

$$U_T = U_A + U_R.$$

If our preferences for achievement scores are dependent on the student's residual score or vice versa, then U_T must be estimated in a more complicated way, by asking lottery questions among many possible achievement and residual score combinations.[2]

18. Question (c). Suppose U_{Ti} has been constructed for each student i. How can U_{Ti} be summed to obtain a school index of success? Once again the answer depends on the interdependence of the components to be combined. If U_{Tk} (the utility for student k) is held fixed at some level $(U_{Tk})_0$, do our conditional (probabilistic) preferences for any other U_{Ti} depend on what fixed level $(UT_k)_0$ is chosen? If not, and if the question can also be answered negatively for all U_{Ti} fixed, then U_{Ti} for all students 1,..., n are mutually preferentially independent.[3] If this independence holds, then U_T(school) can be expressed as an additive value function:

---

[1] Raiffa (1969).

[2] See Raiffa (1971) for details; Raiffa (1968), Ch. 9, Sec. 3, for an outline of the complexities.

[3] Mutual preferential independence means that the decisionmaker's substitution rate between U_{Ti} and U_{Tj} does not depend on any of the values of components other than i and j. See Raiffa (1971), pp. 74-75.

$$U_T(\text{school}) = U_{T1} + U_{T2} + \ldots + U_{Tn}.$$

In other words, if mutual preferential independence exists, evaluating a school merely involves evaluating each student and summing up the utilities over all students in the school.

Unfortunately from the point of view of analytical simplicity, such independence seems not to hold across students. As soon as distributional considerations enter--when we care about equality of outcomes, for example--then our feelings about U_{Tk} *do* depend on the levels of the other students. Furthermore, if part of the education's value is as a screening or credentialling device, then each student's scores affect the utility of his comrades' scores. Therefore, mutual preferential independence does not seem to exist. As a result, $U_T(\text{school})$ can be assessed only through a very complicated series of tradeoffs, holding each U_{Ti} fixed at different levels while assessing the remaining $U_{T(n-1)}$: a theoretically possible but operationally unpalatable task.

19. Using the school mean score as the evaluative statistic assumes a linear utility function and mutual preferential independence, neither of which seems true.

FROM THEORY TO PRACTICE IN EDUCATION

20. How far have we come? We have seen that relying on the mean for evaluation entails assumptions about educational objectives that probably are untenable. We have examined a methodology for determining precisely how to go beyond the mean. This methodology had three parts: eliciting utility functions for achievement and for residual scores for individual students, combining these two functions into a single U_T for each student, and then combining U_{Ti} for each student i into the appropriate evaluation measure for a school. These three steps might be very complex, but theoretically, a rational decisionmaker could produce precisely the "statistics" needed for educational evaluation.

21. Turning from theory to reality, however, two important facts about education must be reckoned with:

(1) Local school districts (and, within districts, various interested parties) are likely to have different utility functions.

(2) Practically, it will be extremely difficult to obtain an operational specification of utility functions from educational decisionmakers.

These two propositions have serious implications for educational evaluation. Both make the methodology of utility functions less than perfectly applicable.

22. The first point implies that the search for a national objective function that somehow *combines* local preferences is futile. Consensus on education objectives will not be forthcoming--and perhaps rightly so. In a decentralized educational system, local preferences possess a certain autonomy, a certain right to be different. To evaluate all schools by the same criteria, with the same utility function, would be an error.[1]

23. The second point means that, in educational evaluation, the objective is not specified in advance. The problem, in my opinion, is not that objective functions are theoretically impossible to get: Roche (1971) has obtained utility functions for raw achievement from

---

[1] Note that the current ways of using many statistical methods to evaluate schools assume common objective functions (and production functions) among schools. Insofar as schools are trying to do different things, regression coefficients relating certain inputs to a common output may be misleading; coefficients of multiple correlation may be looking at the wrong type of variability; good schools may merely be the ones that are trying to do what one is trying to measure. Even if schools share a common objective, they will probably weight it differently in their tradeoffs among their other goals.

There still may be a justification for making evaluations according to a single objective function. Suppose, for example, that the evaluator is the federal government. A decentralized educational system does not preclude the existence of national-level spillover effects from schooling. The federal government would want to affect the local production of these effects through grants-in-aid, legal constraints, taxes, and so forth, even if not through overt control; and the federal government could evaluate its success at doing so with a single national-level objective function that gave utility to the particular spillovers in question. This would, of course, be a very limited sort of evaluation, but perhaps this is all the federal government ought to attempt in a decentralized system.

a school district in New Jersey, following the generally applicable methodology of Raiffa and others. The constraint is instead one of feasibility. Three problems may be mentioned: cost; the ticklish task of defining decisionmakers among the many educational officials with interests and pretensions; and if there are multiple decisionmakers, combining their objectives in a meaningful way. In practice one cannot begin with tightly defined objective functions and then deduce from them the appropriate way to use achievement measures for evaluation.

24. There are, then, two levels of problems in obtaining the well-specified objectives that are theoretically necessary: between-school disagreement and within-school difficulty of specification. If goals are not agreed upon--and indeed if the differing opinions are not even clearly specified--then evaluation cannot mechanically compare results with objectives according to some well-defined statistics. In spite of the systems analyst's predilections, one cannot worry first about objectives and then about the appropriate measures and statistics of progress. Public education does not submit gracefully to this sort of deductive approach, however methodologically alluring.

25. From the systems analyst's point of view, education is the worst of worlds. First, there are no well-specified objectives and they probably cannot be obtained. Second, evaluations must nonetheless be made. Third, the data is mostly restricted to achievement scores. And finally, most existing large-scale evaluations and governmental data banks use only mean scores. We know *something* about educational objectives--not a sufficient amount to draw curves and derive combinatorial rules, but enough to know that the present reliance on the mean is inadequate.

26. What should be done?[1] The situation is somewhat analogous to the

---

[1]Our systems analyst, an ideal type who nonetheless sometimes speaks with the same voice as more reasonable people we know, might suggest the following: "Since your decisionmakers are diverse and no mathematical algorithm can be conveniently adduced for any one or all of them, why not solve your 'statistics for evaluation' problem by giving the entire distribution of scores for each school to all the decisionmakers? Let them make up their own minds what is important." Visions of policymakers trying to examine hundreds of histograms, or having to compute residual measures according to their individual

one faced in evaluating a nation's economic welfare. Clearly the average income statistic is not enough; clearly, too, no social welfare function has been derived from which the appropriate statistics for evaluation could be deduced. In a democracy, such a function may even be impossible to agree upon (Arrow, 1951). But there is a notable difference. Unlike education, national economic policy *has* employed statistics that go beyond the mean: measures of income distribution, the poverty line, and others. These statistics were not deduced from an objective function, and there is no one set of them that commands universal assent as the best and most efficient. But a number of useful statistics *have* been proposed to measure certain ill-defined although meaningful goals of economic policy. Rather than staying where we are in educational evaluation, or throwing out achievement tests altogether, perhaps we would do well to follow that example.

---

perceptions of the proper control variables, may not occur to our analyst (or, if they do, they may only cause him glee). We do not want to overwhelm decisionmakers with data. A map on a 1:1 scale is of little use. Our goal is to provide a few easily computable, informative statistics that correspond (roughly) to the policymaker's likely educational objectives and that are likely to increase his knowledge of educational outcomes.

IV. MEASURES FOR EDUCATIONAL EVALUATION

MEASURES OF CENTRAL TENDENCY

27. Despite the disparaging remarks about current use of the mean, for many purposes an indicator of a school's central tendency is of the utmost importance. The school's average achievement score is a good proxy for the general level of absolute cognitive performance among its students. Similarly, the school's average residual score may be a useful indicator of the general level of cognitive performance relative to nonschool background factors--the school's "value-added" to student cognitive performance compared with other schools having similar students.

28. But is the mean the best estimator of a school's central tendency? F. R. Hampel writes:

> As everybody familiar with robust estimation knows, the mean is a horrible estimator, except under the strict normal distribution; it gets rapidly worse even under very mild deviations from normality, and at some distance from the Gaussian distribution it is totally disastrous.[1]

The mean is greatly affected by observations away from a center of the distribution. If repeated medium-sized samples (say, N = 20 to 50) are drawn from a distribution with a number of such observations--or, as Hampel says, from a distribution that does not closely resemble the normal--the sample means will vary quite markedly.

Under such conditions, a more "robust" estimator is desirable. One with which everyone is familiar is the median. Even the median has its difficulties as an estimator, however, and a recent research seminar has posited and tested more exotic and efficient statistics ("hubers," "sitsteps," folded medians, and others) (Andrews *et al.*, 1972). There was no "best estimator" for all cases: The preferred statistic depends on the sample size and the degree of "contamination" in the underlying distribution.

---

[1] Andrews *et al.* (1972), p. 243. See also pp. 239-240.

29. There is another consideration. If evaluations utilize residual scores, the choice of the appropriate measure of the central tendency can affect the goodness of the fit to student background factors; the choice of measures might depend on the choice of fitting techniques.[1]

30. In part, the choice of estimators depends on pure convenience. The mean and median are well-understood, readily computable measures. The larger the unit of analysis (states, districts, schools, grades in schools, etc.), the less it matters which measure is chosen, and therefore the greater the tendency to use the familiar mean. If N is relatively small (5 to 20), the median is probably a better simple choice.

31. Another problem arises from the use of residual scores for aggregates of students. If the regression against student background factors is made at the student level and then individual student residuals are aggregated to obtain a school residual average, one number is obtained. If, however, the easier fit of *school* mean achievement to *school* background factors is performed, a different number will usually result. Dyer, Linn, and Patton (1969) found that the correlation between such measures was "unsatisfactorily low." Again, how one calculates the central tendency measure can affect a school's score considerably.

32. Neither the school mean nor the residual mean, regardless of how computed, will be a perfect indicator of a school's true "central tendency." But perfection is not the goal; there is not a precisely defined objective. It is enough that both sorts of central tendency measures are useful indicators of the *general* level of performance within a school.

33. Both school means and school residual means turn out to be distributed fairly normally across schools. For large numbers of schools, the Gaussian shape of the unadjusted achievement scores is, of course, expected, since the student scores are normalized. The residual school score finding is more surprising. No matter how simple or complex their control variables, and across many data sets, grades, and years,

---

[1] Andrews *et al.* (1972), p. 131. See also Bickel (1973). Most fitting methods require the mean; for new techniques that do not, see Tukey (1970).

Klitgaard and Hall (1973) found symmetric distributions with thin tails and no evidence of discontinuities across schools. A typical histogram of school mean residual scores is shown in Fig. 2.

Residual means for the same schools do not correlate very highly over time (Jencks, 1972; Acland, 1972; Forsyth, 1973), ranging from 0.05 to 0.50 depending on the test, the years, and the data base. This low correlation might result from random variation or statistical error or else from the fact that school capabilities really do fluctuate from year to year. In a test of the statistic's consistency, Dyer, Linn, and Patton (1969) correlated mean residuals for random halves of the same school class in the same year and found a correlation of 0.88 for the composite test. The lack of a perfect correlation results, of course, from imperfectly matched samples, test error, and so on, as well as from shortcomings in the residual mean statistic itself. In general, this relatively high correlation is encouraging for potential users of residual scores in evaluation.

The median correlation between the uncontrolled school mean achievement score and the residual score[1] over eight sets of Michigan elementary school data in 1969-1970 and 1970-1971 was 0.55. Of course, this less-than-perfect correlation reveals that background factors do explain much of the variation in achievement scores.

SPREADS

34. Equality is an increasingly voiced goal of education. In America discussions of equality have traditionally centered on equality of opportunity: that everyone have an equal *chance* to obtain a good education, but not necessarily that everyone actually use that chance. However, many recent writers, including some of a radical bent, have emphasized equality of *outcomes* as a major educational aim. They maintain that instead of evaluating some prior notion of the opportunity schools provide--or perhaps in addition to such an investigation--the equality of the actual results should be examined.

---

[1]The regression equations are given in Klitgaard and Hall (1973), p. 46.

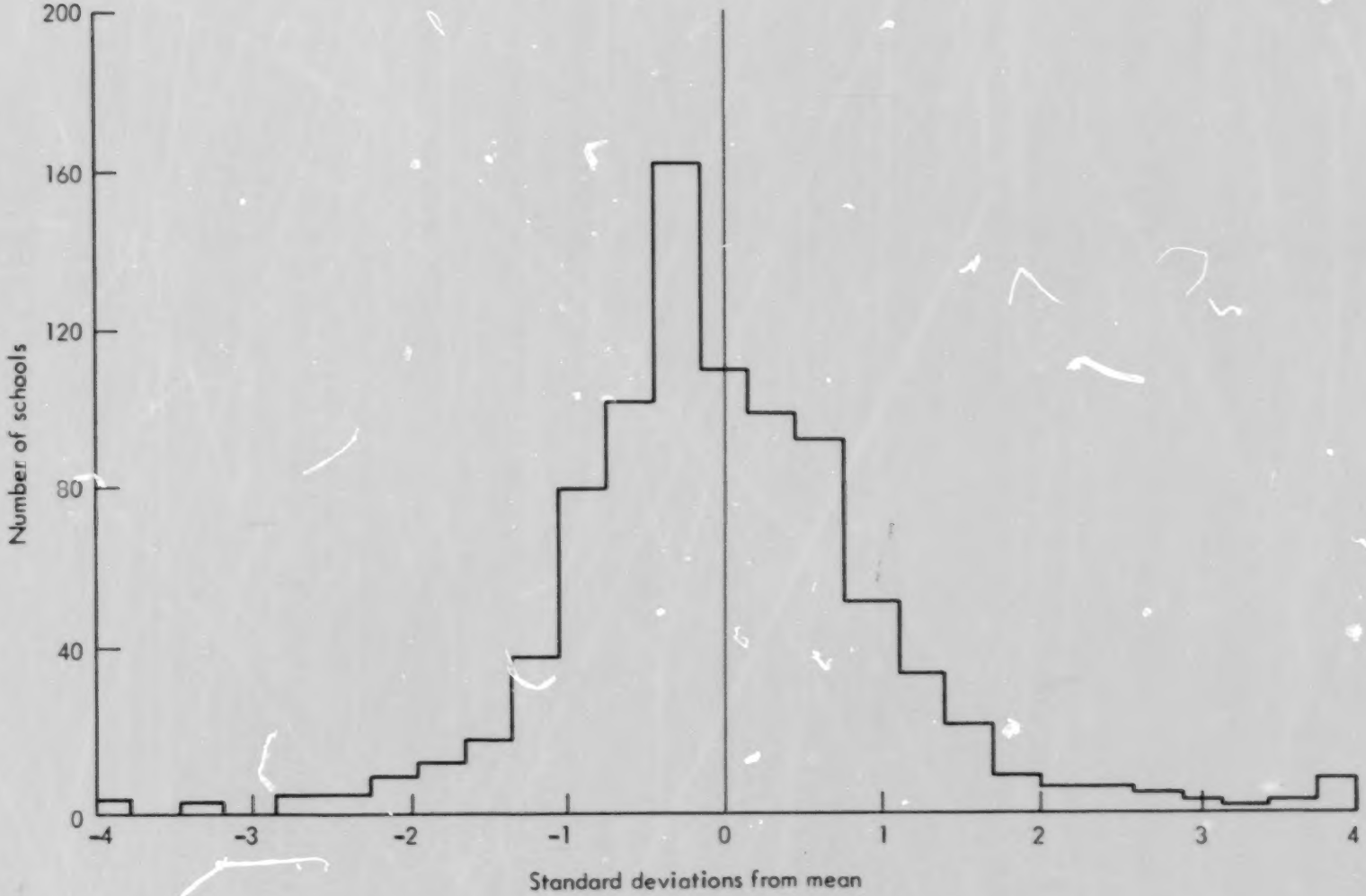


Fig. 2 — Histogram of residuals for 1970–1971 Michigan seventh-grade mathematics test, from a regression controlling for racial composition, community type, and socioeconomic status (Klitgaard and Hall, 1973, p.36)

It is not clear that the more equal the educational outcomes, the better; one's utility function might not be an increasing function of the amount of equality.[1] The central point is not that equality is preferred indefinitely but that some measure of the equality of outcomes that a school provides is helpful in a well-rounded evaluation of its effectiveness.

A school's mean score alone tells nothing about its equality of outcomes (although a comparison of school means will indicate something about equality among schools). To evaluate a school's equalizing ability, one needs to go beyond its central tendency to some estimator of the spread of the school's distribution of achievement scores.

Figure 3 shows two hypothetical distributions of achievement scores corresponding to schools A and B. Other things equal, an advocate of equality of outcomes would prefer school A because of its smaller variability, even though the school mean scores are equal.

One statistic of interest, then, is the spread of a school's uncontrolled achievement scores. Other things equal, the smaller the spread, the greater the equality of cognitive achievement outcomes.[2]

35. Two kinds of residual scores related to the spread can also be useful. First, suppose one is interested in comparing schools' *equalizing* abilities. The different degrees of equality within schools may stem from differences in nonschool background factors from school to school, rather than different equalizing effects in schools. Schools having students with more similar backgrounds can expect less variation in achievement scores. One could regress some statistic of equality of outcomes (say, the standard deviation of school scores) against various background factors to compute a predicted standard deviation for each level of the background variables. A residual score--observed standard deviation minus predicted standard deviation--could then be obtained for

---

[1] Despite the common usage of terms like "equality" as if they were to be maximized, there is almost surely some limit in everyone's mind--although, as Kristol (1972) points out, advocates of equality and mobility are reluctant to define optimum levels.

[2] Some educators apparently believe that larger spreads indicate superior schooling: "Every experienced teacher knows that effective teaching will increase the variance of the group being taught, and usually markedly" (Guba, 1967, p. 61).

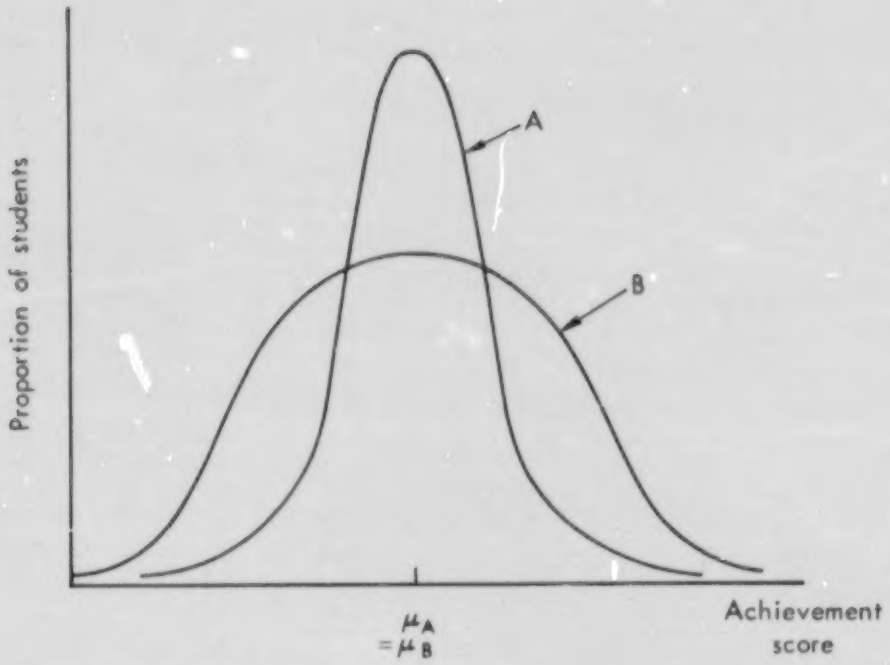


Fig. 3 — Schools with equal means and unequal spreads

each school. The smaller this residual, the greater a school's equalizing ability.

36. A second residual spread measure might serve as a proxy for "educational mobility," another goal of schools. Americans have long cherished the belief that education can be a powerful weapon for social advancement, without students being imprisoned by their socioeconomic backgrounds. Some recent studies, using mean achievement scores, have eroded this faith. But is the mean the right statistic to measure the effects schools have on mobility?

Suppose the mobility objective is not fulfilled by small movements of average school performance but by providing *some* students the chance to succeed far beyond what would be predicted by their socioeconomic backgrounds. Suppose parents would be willing to pay for "good" schools even though only one child in twenty would do much better than in a "bad" school (and some children might do worse). They might feel that a *chance* for high mobility was worth the cost. (In this sense, mobility and educational opportunity are linked as objectives.) In such cases the mean will not be a useful measure of school success.

For this mobility objective, the spread of achievement *residuals* may be a useful indicator. (In general, the spread of the residual scores will not be the same as the spread of the raw scores.) Given schools with equal mean residuals, the one providing greater residual variation is providing greater educational mobility. Their students have more opportunity to "succeed"--*and* more to "fail"--compared with other schools whose students have like socioeconomic and personal characteristics. Putting it another way, the students in a school with a larger variation of residual scores are less likely to end up where their backgrounds would have predicted.

As with equality of outcomes, it is not necessarily true that the more such "opportunity" for success and failure, the better. One may prefer to have less chance of failure even at the loss of some opportunity for success. In 1523 on the Isla de Gallo, Pizarro drew a line with his sword in the sand and told his men that on one side lay "untold hardships and starvation, treacherous reefs and storms, bitter wars and even death, but there also the golden land of the Incas" and

on the other "peace, but the peace of poverty." Only 13 of the hundreds joined him on the side of possible riches. Risk preferences and distributional considerations are important in deciding how much opportunity for mobility we prefer.[1] The fact that mobility may not be indefinitely preferred does not, however, mean that the spread of residual scores is a useless measure. It is merely a reminder that "mobility" is two-directional, and that more of it, in education as elsewhere, may not be unequivocally desired.

37. There are, then, three possible measures of spread that would be useful in educational evaluation: the spread of the unadjusted achievement scores, indicating equality of outcomes; the difference between the actual and expected spread of achievement scores, a proxy for the equalizing ability of schools; and the spread of the residual scores of a school's students, indicating the amount of educational mobility a school provides.[2] Which statistic should be selected to measure spread?

As in the case of income distribution (see p. 5, note 1), there are many possible measures of dispersion and equality. The most common for statistical applications is the variance (or its positive square root, the standard deviation). However, like the mean but even more so, the variance is very sensitive to extreme values; it is not a robust estimator of spread. One estimator of spread that is less vulnerable to outliers is the interquartile range (others are given in Tukey, 1970, Vol. I, Ch. 2).

---

[1] Risk preferences are important because people with higher risk aversion tend to prefer narrower distributions of outcomes to wider ones, given equal expected values.

Distributional considerations may enter if the residuals display heteroscedasticity. In such cases an increase in the overall variance of a school's residuals increases the opportunities for students of certain backgrounds more than others; one cannot *a priori* presume that every student has the same probability of being located anywhere on the school's distribution of residuals. Therefore, *which* students get more opportunity becomes paramount--and this brings distributional objectives into the picture.

[2] The last two measures are obviously similar, but they are not mathematically identical. Unfortunately, my data did not allow me to calculate the third measure, to see how it relates empirically to the second; this is one of many research tasks required to tie down the ideas in this report.

Which statistic to use for evaluation should, as before, depend on a careful specification of the educational objective function; but, short of this, what matters is that *some* such measures of spread be available. Further research should be devoted to selecting the best statistics of spread for education, although as in income distribution, optimality properties may not be agreed upon. With any of a number of measures of dispersion, schools could be compared cross-sectionally and over time in a useful way; the value of such statistics for evaluation should not be underestimated because of some misplaced desire for cardinal precision.

38. How do various measures of spread in educational data behave? How much do schools differ in the spreads of their achievement scores? Do nonschool background factors explain differences between the spreads of schools? Is there any evidence that some schools consistently provide less variability of scores than others, holding nonschool factors constant? Since spread measures of the interschool distribution of test scores have largely been ignored in the past, little is known about the empirical characteristics of such measures.

The following are merely preliminary investigations into the behavior of some standard deviation measures based on Michigan data for fourth and seventh grades in 1969-70 and 1970-71.[1] Since the data were already aggregated at the school level, the "mobility" statistic, which must be based on student-level regressions, could not be computed. Only the standard deviation of unadjusted scores ("equality" statistic) and the difference between the expected and the observed standard deviation ("equalizing ability" statistic) were examined, and these two only in an exploratory fashion.

How should one expect the standard deviation statistic to behave? It is the square root of the variance, and it is similarly sensitive to extreme values in the distribution. In normal samples, the sample variance is distributed as a multiple of a Chi-square variate with N-1 degrees of freedom. With N small (say, less than 10), the Chi-square distribution is positively skewed; but by N = 20, the distribution is

---

[1] The data base is described in Brown (1972) and Klitgaard and Hall (1973).

close to Gaussian (Brownlee, 1965, pp. 82 ff). The standard deviation tends to have higher variability for smaller N; schools with fewer students tested will have a higher proportion of high and especially low standard deviations, other things equal.[1]

In the Michigan data N (the number of students tested per grade) varied considerably from school to school (see Table 2), making school standard deviations not perfectly comparable; but since the average value of N was quite large, the analysis simply used the standard deviation without worrying about transformations. Eliminating all schools with $N < 5$, the average school standard deviation was about 9 and the standard deviation of the standard deviations was about 1.1 (see Table 2).[2] Two distributions of school standard deviations are shown in Figs. 4 and 5. Notice that the distributions are negatively skewed. This interesting fact held throughout all eight distributions.[3] It might reflect the lower variances of smaller schools. It also might indicate that some schools are trying to obtain more equality of outcomes than others, or are better at doing so than other schools with similar goals.

How do these standard deviations compare with those expected, given the different background factors among the students of different

---

[1]The variance of the sample *variance* is $1/n[\mu_4 - (n-3)/(n-1)\sigma^4]$, where μ_4 is the fourth central moment of the population distribution (Wilks, 1962, p. 199). Because the distribution of sample variances is positively skewed for small N, small schools will tend to have more low scores than high ones.

[2]The achievement tests are normed to have an interstudent standard deviation of 10 and mean 50.

[3]The data cover reading and mathematics scores for fourth and seventh grades in 1969-70 and 1970-71, a total of eight sets. Not every school has both fourth and seventh grades, and not every school reported data for each possible test/grade/year combination. The skewness statistics were:

R 4 69-70 = -0.48	R 4 70-71 = -0.62
M 4 69-70 = -0.47	M 4 70-71 = -0.28
R 7 69-70 = -0.68	R 7 70-71 = -0.64
M 7 69-70 = -0.94	M 7 70-71 = -0.98

R 4 69-70 stands for the reading score for fourth grades in 1969-70; the other symbols are interpreted similarly.

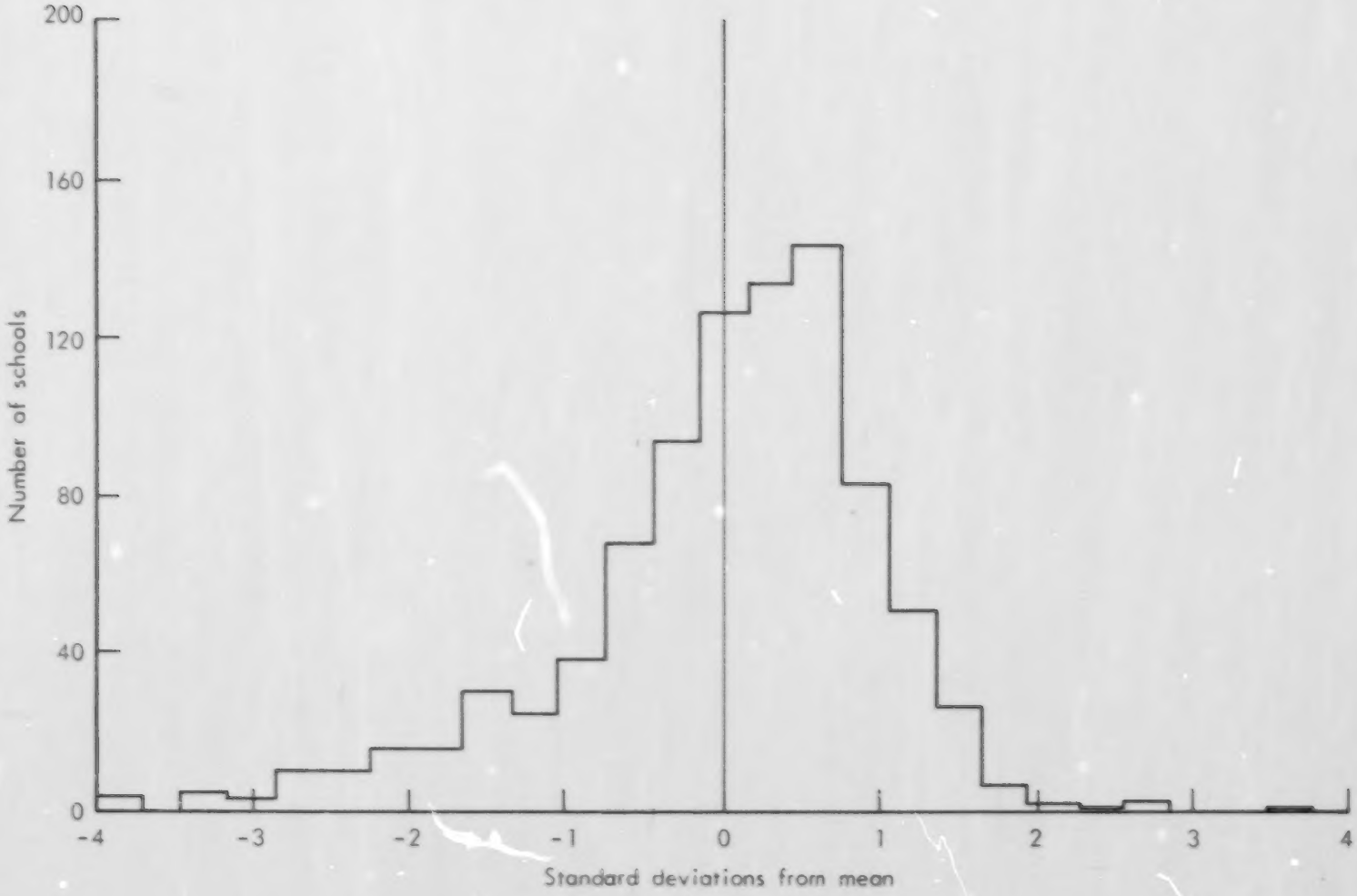


Fig. 4 — Histogram of uncontrolled school standard deviations for 1969-70 Michigan seventh-grade mathematics scores

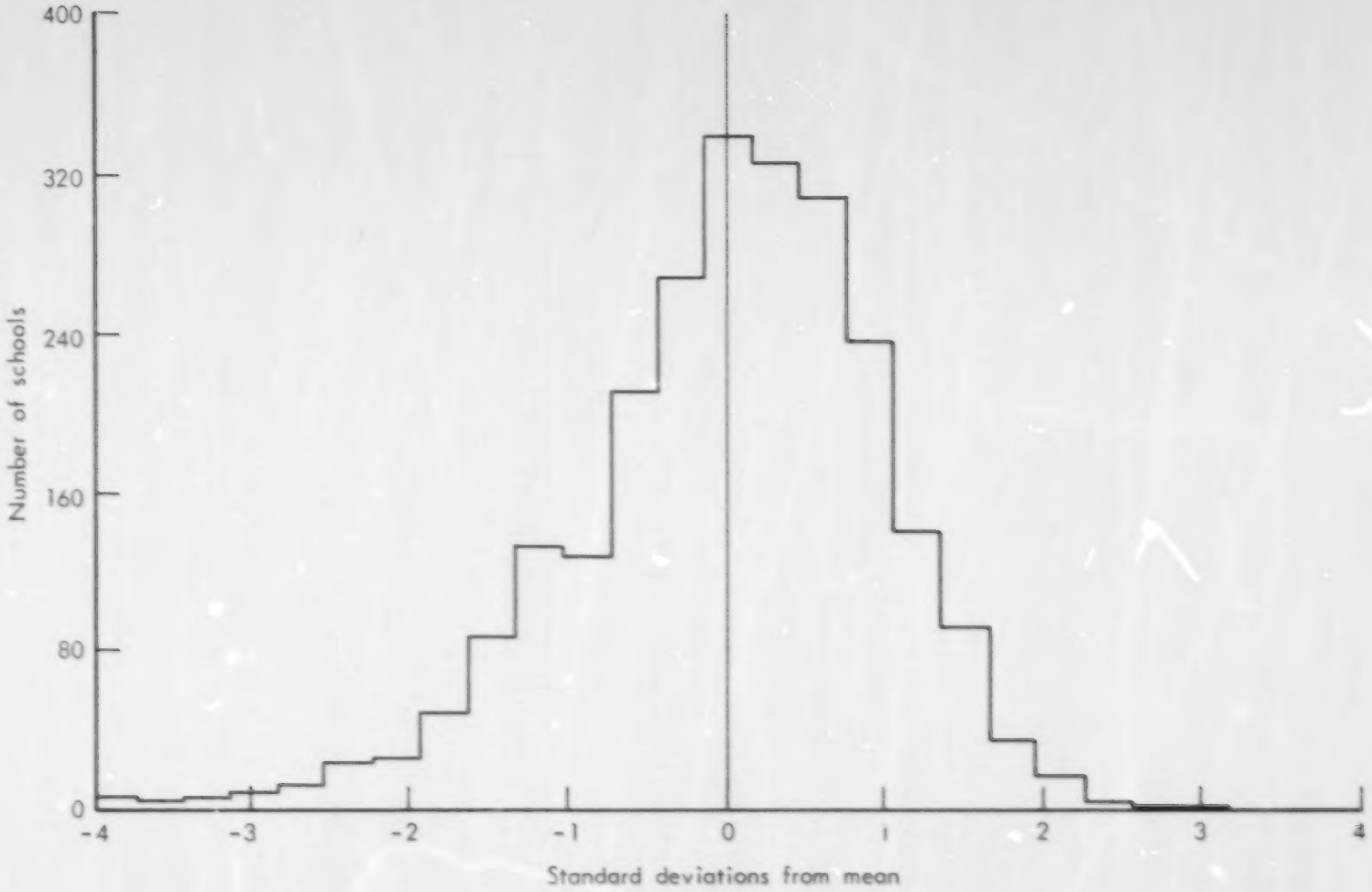


Fig. 5 — Histogram of uncontrolled school standard deviations for 1970-71 Michigan fourth-grade reading scores

schools? To find out, a series of regressions were run, fitting the school standard deviation to a number of nonschool background factors. The best set of regressions, although still only crude and exploratory, is given in Table 1. Table 2 shows the means and standard deviations of the regressor and response variables.

The proportion of variation explained by the regression results varies rather widely, from 0.11 to 0.37. No differences seem important between the reading and mathematics regressions, although the reading scores display more heteroscedasticity as indicated by the greater significance of the $\bar{x}$ regressor.[1] (This difference is most striking between the fourth grade reading and mathematics scores.) SES has the expected positive sign on all regressions. %MIN is consistently negative, indicating that greater numbers of minority students tend to go along with lower standard deviations, even after controlling for SES and the achievement score $\bar{x}$. The number of students tested N has the expected positive sign, indicating that smaller schools do tend to have smaller variability.[2]

The major finding of these regressions and the others that were tried is the limited ability of background factors to predict school standard deviations. This result, of course, contrasts markedly to the results of regressions on school means, where most of the variation across schools is explained by socioeconomic, racial, and regional variables. (For example, the R^2 values for simple regressions on means using the same Michigan data ranged from 0.59 to 0.78 (Klitgaard and Hall, 1973, p. 46).) One might hypothesize that the low explanatory power of background factors indicates that school policies determine standard deviations. But the low R^2 values may merely be a product of

---

[1] Heteroscedasticity refers to nonconstant variance of residuals around the regression line. The significance of the negative mean achievement regressor for the reading scores indicates that the variation of scores was greater in schools with lower average scores.

[2] The statistical properties of the standard deviation statistic would lead one to expect smaller variances for schools with small N, even if all schools had drawn their students randomly from the same population. It also may be true that smaller schools tend to have more homogeneous student bodies, even after controlling for SES.

Table 1

MICHIGAN SCHOOL REGRESSION ON STANDARD DEVIATIONS OF TEST SCORES

Test	Equation	R^2	Std. Error	F	Number of Schools
R469-70	$\sigma = 17.8 + 0.14\ (SES\sigma) - 0.20(\mu) - 0.03\ (\%MIN) + 0.003(N)$ (72.5) (764.4) (458.8) (25.6)	0.26	1.03	206.9	2376
M469-70	$\sigma = 7.7 + 0.12\ (SES\sigma) + 0.003(\mu) - 0.01\ (\%MIN) + 0.004(N)$ (49.8) (0.2) (156.7) (35.1)	0.13	1.04	86.1	2376
R769-70	$\sigma = 9.7 + 0.17\ (SES\sigma) - 0.04(\mu) - 0.01\ (\%MIN) + 0.001(N)$ (35.0) (9.5) (38.0) (10.7)	0.11	0.88	25.5	870
M769-70	$\sigma = 8.1 + 0.19\ (SES\sigma) - 0.01(\mu) - 0.02\ (\%MIN) + 0.001(N)$ (40.9) (1.2) (125.4) (19.7)	0.25	0.92	71.5	870
R470-71	$\sigma = 19.9 + 0.08\ (SES\sigma) - 0.24(\mu) - 0.03\ (\%MIN) + 0.006(N)$ (28.6) (1147.4) (453.2) (79.4)	0.37	1.03	345.0	2401
M470-71	$\sigma = 8.5 + 0.10\ (SES\sigma) - 0.01(\mu) - 0.02\ (\%MIN) + 0.004(N)$ (41.8) (2.6) (188.7) (36.2)	0.32	1.07	67.3	2401
R770-71	$\sigma = 15.0 + 0.17\ (SES\sigma) - 0.14(\mu) - 0.02\ (\%MIN) + 0.001(N)$ (54.6) (140.8) (123.2) (14.4)	0.23	0.86	60.9	841
M770-71	$\sigma = 6.2 + 0.13\ (SES\sigma) + 0.03(\mu) - 0.02\ (\%MIN) + 0.001(N)$ (33.2) (7.0) (80.4) (30.3)	0.22	0.87	59.1	841

R469-70 stands for the reading scores for the fourth grades in 1969-70. The other symbols are interpreted similarly (M = mathematics). σ = test score standard deviation; $SES\sigma$ = socio-economic status standard deviation; μ = test score mean; %MIN = % pupils of minority races; N = number of students tested. Figures beneath the regression coefficients are the F-ratios ($=t^2$).

Table 2

MEANS AND STANDARD DEVIATIONS OF REGRESSOR AND RESPONSE VARIABLES

	μ	σ		μ	σ
R469-70μ	50.5	4.0	SES 469-70 σ	8.8	1.4
σ	8.9	1.2	SES 769-70 σ	8.6	1.2
N	62.8	34.7	SES 470-71 σ	8.8	1.4
M469-70μ	50.5	4.0	SES 770-71 σ	8.8	1.4
σ	9.0	1.1	%MIN 469-70	10.7	23.7
N	62.8	34.7	%MIN 769-70	10.5	22.6
R769-70μ	50.3	3.2	%MIN 470-71	10.1	22.8
σ	9.2	0.9	%MIN 770-71	9.4	21.1
N	172.7	130.2			
M769-70μ	50.4	3.8			
σ	9.0	1.1			
N	172.4	129.8			
R470-71μ	50.6	3.9			
σ	8.9	1.3			
N	63.5	34.4			
M470-71μ	50.6	4.2			
σ	8.9	1.1			
N	63.4	34.3			
R770-71μ	50.6	3.3			
σ	9.2	1.0			
N	182.5	133.9			
M770-71μ	50.6	3.9			
σ	9.0	1.0			
N	182.1	133.3			

greater random fluctuation or purely statistical problems. This question awaits detailed investigation.

The residuals from these regressions constitute the second spread measure discussed above--a statistic purporting to indicate the equalizing ability of schools given their students' backgrounds. Figures 6 and 7 show the distribution of residuals for the same two tests as Figs. 4 and 5. The distributions are slightly tighter: The standard deviations (of the standard deviations) now average about 1.0. Skewness has been reduced, although all eight distributions are still negatively skewed.[1] Outliers remain on the left tail, but a few also show up on the right tail now.

The extreme values on the left tail looked interesting enough to pursue. Each histogram of schools' scores (say, for a particular grade, test, and year) will show the effects of random variation as well as the effect of different schools. A thick left tail does not by itself prove that these schools with low variability are anything more than random deviates. But if the same schools show up on the left tail consistently over many grades, tests, and years, one might conclude that the phenomenon is not just a statistical fluke. Do some schools consistently record low variability, even after allowing for nonschool background factors?

To find out, the following null hypothesis was formulated: All variation of the difference between actual and expected standard deviations is a result of chance and not of school effectiveness. To test this hypothesis, some sort of "cumulative distribution" is required indicating how well schools have done over many grades, tests, and years after controlling for background factors. Then it would be possible to see if that

---

[1] Skewness statistics were:

R 4 69-70 = -0.20	R 4 70-71 = -0.26
M 4 69-70 = -0.14	M 4 70-71 = -0.01
R 7 69-70 = -0.36	R 7 70-71 = -0.37
M 7 69-70 = -0.16	M 7 70-71 = -0.19

Notice especially how the mathematics scores have become less skewed.

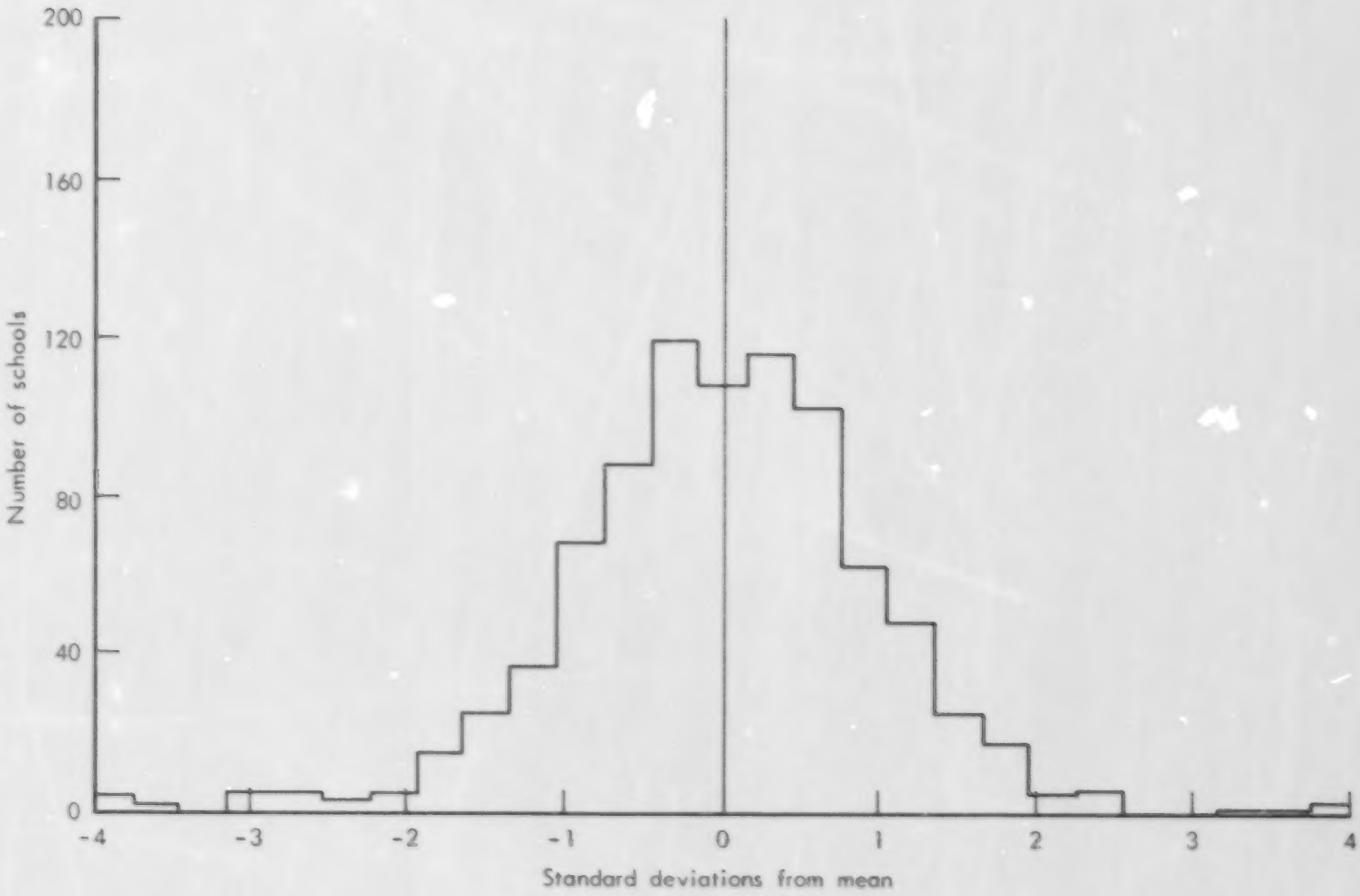


Fig. 6 — Histogram of residuals from a regression of school standard deviations for 1969-70 Michigan seventh-grade mathematics score against socioeconomic status standard deviation, mean seventh-grade mathematics score, percent minority students, and number tested

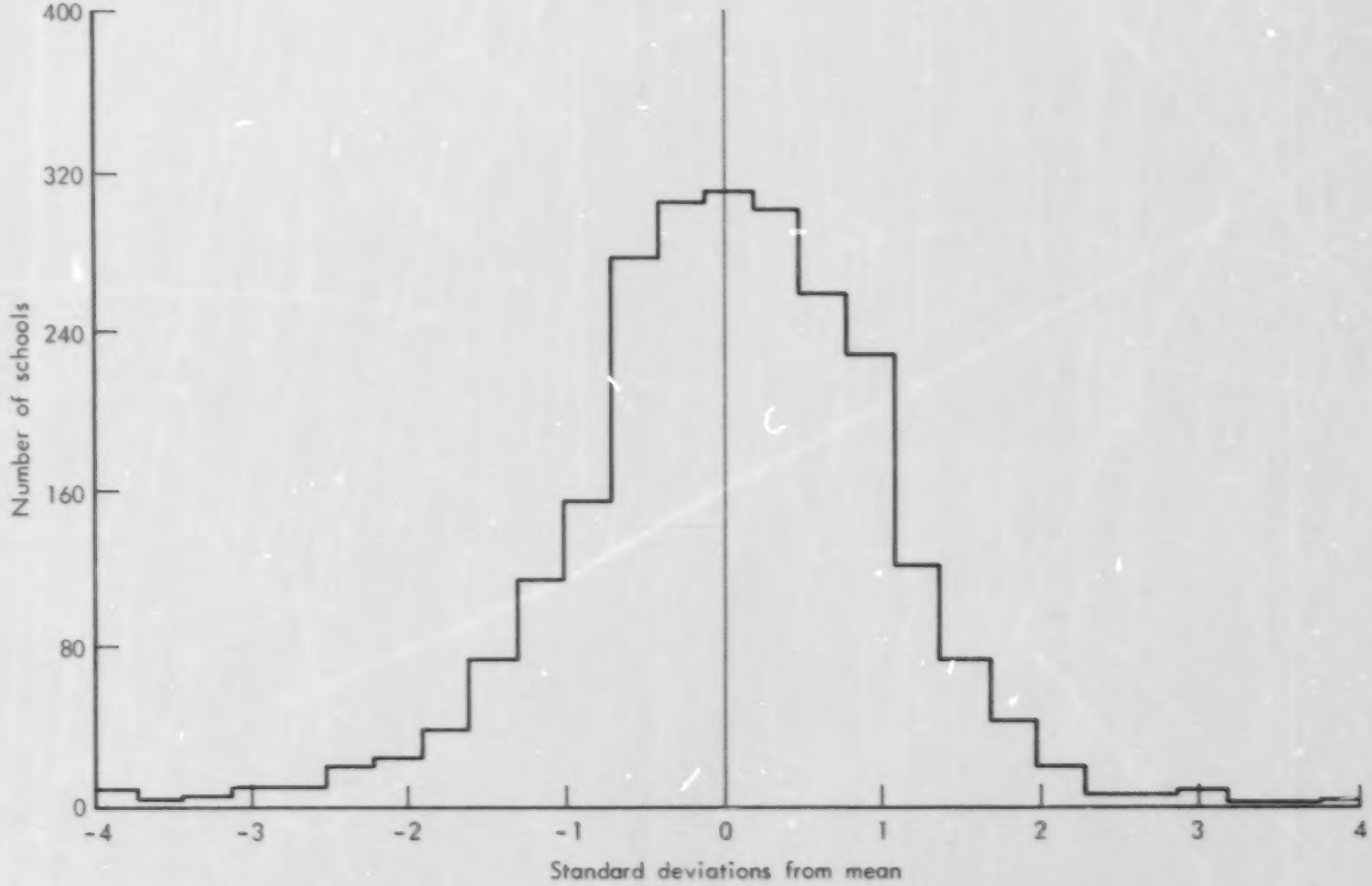


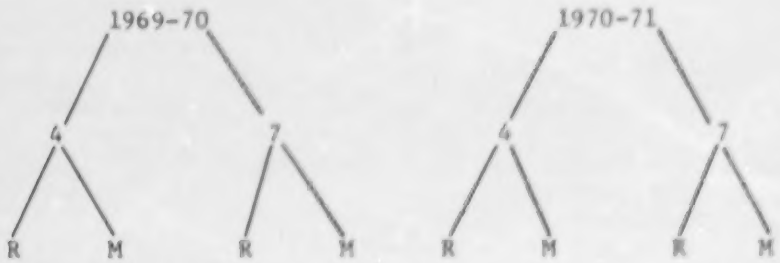
Fig. 7 — Histogram of residuals from a regression of school standard deviations for 1970-71 Michigan fourth-grade reading score against socioeconomic status standard deviation, mean fourth-grade reading score, percent minority students, and number tested

distribution differed significantly from a theoretical distribution obtained by treating all the individual distributions of residuals as statistically independent.

As a proxy for this cumulative distribution, each school in a given grade, test, and year distribution was assigned a score of one if it was more than one standard error below the mean and a score of zero otherwise. Each school's totals were added up over all distributions, and a Chi-square test was used to see whether some schools were consistently below one standard error more than chance would predict. The results appear in Table 3.[1]

---

[1] A deviation from the assumption of perfect independence of the various test scores was necessary to take account of the correlation between reading and mathematics residuals in tests taken by the same class in the same year. The tree below shows how the eight residuals were generated:



Since the R-M residuals for a given year and grade are not independent, the null hypothesis was reworded to posit that the pairs of scores are independent.

Let X_i be the number of scores in a school's reading-mathematics pair (R_i, M_i) that are one standard deviation below the mean. X has the possible values 0, 1, 2. Now compute a total score T_j for each school where $T_j = X_1 + X_2 \ldots + X_j$ (j is the number of pairs of scores the school reported). Assuming the X_i are independent, compute null distributions for T_j using the actual probabilities of 0, 1, and 2 successes per pair. Then the actual distribution can be compared with the null distribution using a Chi-square test.

The actual probabilities for each pair of tests are:

Table 3

RESULTS OF CHI-SQUARE TESTS OF DIFFERENCES BETWEEN
OBSERVED AND EXPECTED DISTRIBUTIONS OF RESIDUALS

Schools Reporting 8 Times				Schools Reporting 6 Times				Schools Reporting 4 Times		
No. <-1σ	Observed		Expected	No. <-1σ	Observed		Expected	No. <-1σ	Observed	Expected
0	52		63	0	61		75	0	1852	1742
1	26		40	1	33		36	1	484	552
2	23		25	2	20		20	2	186	257
3	13		8	3	17			3	49	34
4	15			4	4	23	6	4	21	6
5	4			5	2					
6	2	25	3	6	0					
7	2									
8	2									
Chi-square = 167.0				Chi-square = 49.0				Chi-square = 74.3		
Degrees of freedom = 4				Degrees of freedom = 3				Degrees of freedom = 4		

All residuals were derived from a fit of the achievement score standard deviation against SES standard deviation, achievement score mean, percent minority enrollment, and number of students tested. All Chi-square statistics are significant beyond the 0.005 level.

There is evidence in Table 3 that some schools consistently have a greater equalizing effect on their students' achievement scores than chance alone would predict. The causes for such phenomena could not be assessed within the scope of this report.

The schools that were consistently below average did tend to be quite a distance below each time. For example, the ten schools that were below one standard deviation at least five out of eight times averaged about 1.6σ below each time. Since the standard errors were about one test score point and the interstudent σ = 10, these ten schools were reducing the variability of their students' scores about 1/6 of the interstudent variation compared with the average school. On the fourth grade Iowa reading test, this would imply tightening the standard deviation of outcomes about 20-25 percent of a grade equivalent.[1]

(R_i, M_i)	P(X=0)	P(X=1)	P(X=2)
4 69-70	0.802	0.149	0.049
7 69-70	0.823	0.124	0.053
4 70-71	0.804	0.139	0.057
7 70-71	0.845	0.102	0.052

If the school reported eight scores, it had eight chances to be below one standard deviation less than the mean; the null hypothesis is computed for four pairs of tests. If a school only had six chances, then the test is computed from three pairs; if four chances, two pairs. The chances only occurred in reading-mathematics pairs (any school that reported a reading score for a given grade and year also reported a mathematics score for that grade and year). For simplicity in calculation, I assumed a common probability distribution P(X=0) = 0.82, P(X=1) = 0.13, P(X=2) = 0.05 for all pairs and assumed it did not matter which particular pairs happened to make up a school's set of chances.

For the Chi-square approximation to be accurate in contingency tables with more than one degree of freedom, cells with small expectations must be pooled. I followed a pooling rule proposed by Yarnold (1970, p. 865):

> If the number of classes s is three or more, and if r denotes the number of expectations less than five, then the minimum expectation may be as small as $5r/s$.

[1] Lindquist and Hieronymus (1964). To give another intuitive idea of what this reduction in variability means, a 1/6 reduction in the standard deviation on most IQ tests would be 2-3 points.

A final note: The median correlation between the unadjusted standard deviation and the residual was 0.85 across the eight tests.

It must be reemphasized that these results are only explorations. They have barely touched the surface of the important questions concerning standard deviation and other spread measures in education. How do different measures of spread behave? How important is the variability involved? How does spread relate to school and background characteristics? Perhaps this beginning can whet some appetites and suggest some directions for further study.

DISTORTIONS

39. In recent years especially, educational policy has laid heavy stress on special programs for disadvantaged and gifted students. Spurred by the conviction that curricula and methods designed for the average pupil do not teach slow and fast learners efficiently, reformers have created programs for special students at an unprecedented rate. Evidently, many educators base their judgments of school quality partly on the number and sophistication of programs for different kinds of students. If educational policy is significantly directed at slow or fast students, a school's average achievement scores may be a misleading measure of its success.

40. Take the case of uncontrolled achievement scores. Suppose very low scores are very undesirable, very high ones extremely nice, and those around the middle more or less the same. Low achievers might be harmful to society to a far greater extent than the linear weighting of their achievement scores would indicate, while high achievers might be deemed extremely valuable. In this case, the utility function might look like the cubic function in Fig. 1c. We may be willing to let those in the middle achievement range drop a little if we can thereby move both tails of the distribution of scores to the right. For example, in Fig. 8 we may prefer school A to B, and either A or B to school C, despite equal means and variances. Distribution A has more students below the mean than B, but most are in the range where it does not matter too much; meanwhile, A's lower tail is smaller and its upper tail broader.

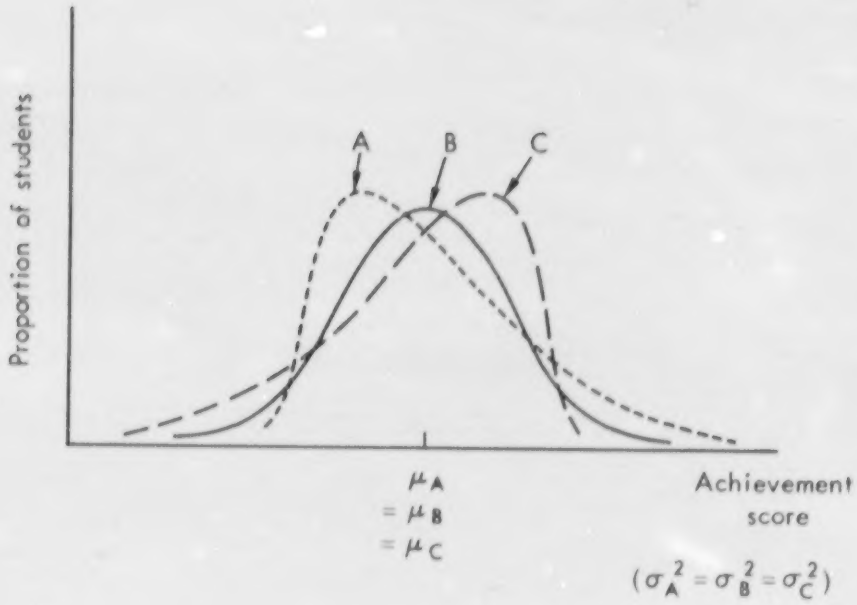


Fig. 8 — Schools with equal means and variances but unequal skewness

One proxy for such preferences might be the skewness of the distribution, defined as

$$E\left(\frac{(X - \mu)^3}{\sigma^3}\right).$$

Positive (negative) skewness indicates that for any specified mean and variance, the mode is likely to be smaller (larger) than the mean, the left tail "unusually" short (long), and the right tail "unusually" long (short). Increasing the positive skewness of a school's distribution of scores trades off losses around the middle of the distribution for gains in scores on both tails. Other things equal, much of educational policy probably favors positive skewness.

41. Similar remarks apply to the skewness of the school's distribution of *residuals*. Fit individual student scores against their nonschool background variables; compute individual residuals for each student; then aggregate those residuals by school and compute the school's skewness statistic for the distribution of residuals. Suppose that we care more about underachievers and overachievers (no matter what the score their background factors would predict). If we wish to avoid large underachievers and produce large overachievers, and if we do not care much about performances relatively near to expectation, then, other things equal, the skewness of the distribution of residuals may validly order schools according to our preferences.

42. Because the skewness statistic is a nonlinear functional, strictly speaking there is no von Neumann-Morgenstern utility function corresponding to its maximization. However, despite this rather ungainly feature, the skewness statistic has a history of use in econometric studies to measure exactly the phenomena relevant here: high emphasis on large positive payoffs and great displeasure at large negative ones (Tintner, 1942; Hicks, 1950; Arditti, 1967; Fisher and Hall, 1969). Used in regressions that also control for mean and variance, the skewness statistic--or some such measure of the distortion of the intraschool distribution--seems an appropriate additional measure for educational evaluation.

Once again, the precise mathematical definition of the distortion parameter to be included is not of prime importance, nor would one prefer positive skewness indefinitely. What matters is that some indicator of distortion be available as an evaluative tool. The skewness statistic is extremely sensitive to outlying values--more than the variance or the mean--and a more robust estimator might be called for (for instance, a trimmed skewness estimator, discarding a small percentage of upper and lower values). Another problem concerns the fact that one's preferences for skewness cannot be separated from one's preferences for mean and variance. Even to find a function that ranks distributions in the same order as maximizing the third moment of a distribution $E(X - \mu)^3$ involves specifying the mean and variance as well.[1] With some such measure one can obtain further information that generally goes beyond the mean (which weighs all gains and losses the same no matter where on the distribution they fall) and the spread (which evaluates bigger tails on either end the same).

44. Other things equal, the more positively skewed the distribution of raw scores within a school, the better a school is doing with its slow and fast learners, although at the expense of its average students. And for individual residuals, with other things equal, the more positively skewed the distribution within a school, the better a school is doing with its under- and overachievers, although at the expense of students who perform at about the level predicted by their socioeconomic backgrounds.

PROPORTIONS ABOVE CERTAIN THRESHOLDS

45. If some minimum level of attainment is of concern, the mean school score can easily mislead. A simple and useful measure is available: the proportion of students who score above the level in question.

---

[1] Note that this fact implies a lack of preferential independence among the goals relating to mean, variance, and skewness of a school's distribution: How much skewness one prefers has to depend on the level of the school's mean and variance.

46. A number of writers imply that certain thresholds of achievement are of the utmost concern.[1] Threshold definitions of success are often used in education. High schools are sometimes judged by the proportion of their graduates that can read at the ninth-grade level or that go on to college, to name two quite different thresholds. In performance contracting experiments, fees often depend on the number of students performing at or above their grade levels. For such objectives, the proportion of students above a certain score is the best indicator of success.

47. As with the other statistics discussed so far, the proportion above certain thresholds has useful applications with both uncontrolled and residual scores. The proportion of students above some absolute level tells us one thing about a school; the proportion achieving above some level relative to their backgrounds, quite another. Both measures usually go beyond the information provided by means, variances, and skewness.[2]

48. Some crude indications of how threshold measures behave can be gathered from data from the Yardstick Project in Cleveland, Ohio. Yardstick contracts its data analysis services to some 34 school districts in Ohio and other states. Its clientele varies from year to year, as do the clients' data requests: Some ask for analyses of lower elementary grades and some upper, and over varying time spans. Thus the data base is not necessarily representative nor is it useful for longitudinal

---

[1] A lower tail threshold is implicit in the writings of Kenneth Clark, for example. Similar sentiments may be discerned in the writings of John Stuart Mill:

"It may be asserted without scruple, that the aim of all intellectual training for the mass of the people should be to cultivate common sense; to qualify them for forming a sound practical judgment of the circumstances by which they are surrounded. Whatever, in the intellectual department, can be superadded to this, is chiefly ornamental." (*The Principles of Political Economy*, Book II, Chapter XII: cited in Vaizey, 1962, p. 20.)

[2] If, however, a school's distribution of scores is perfectly normal, the mean, variance, and number of students tested provides a complete description of the distribution. It adds no information in such cases to say what proportion of students scored higher than some particular threshold.

analyses. However, the Yardstick data bank stratifies school data in interesting ways. For instance, it provides growth-per-year scores stratified by five IQ levels and five categories of father's occupation.

For 72 schools separate regressions were run on school mean growth (mean score for year N minus mean score for year N-1), school mean growth for students with IQs higher than 123, and school mean growth for students with IQs lower than 93. Control variables included father's occupation and mean school IQ, among others.

For the school means, a stepwise regression yielded $R^2 = 0.55$. The other fits were very poor. In the regression on school mean growth among its students with IQ > 123, only the percentage of children in the school whose fathers were skilled workers was significant (with a negative coefficient), and the R^2 was only 0.18. On the under 93 side, no variables reached the $F \geq 4$ significance level needed to enter the regression; and when all controls were forced into the fit, the R^2 rose only to 0.13. These results suggest that school variables may make more difference than background factors in determining the achievement of exceptional children, either because schools concentrate their efforts there or because schooling with uniform emphasis across children affects some children more than others.

V. PRACTICAL CONSIDERATIONS AND CONCLUSIONS

To restate the problem: Large-scale educational evaluations and government data systems are throwing away useful information. This problem is not severe with intensive, small-scale studies; they have the time and resources to do thorough data analysis. But large-scale surveys, proposed "accountability" systems, and government information banks rely almost exclusively on mean scores--be the unit of analysis schools, programs, districts, or whatever. There is inefficient program evaluation and managerial use of data.

49. Given this situation and the continual need for policy decisions, there are three undesirable alternatives. First, one can forgo achievement data altogether, relying instead on less quantitative evaluative criteria. Second, one can choose to remain with mean scores alone. Third, one may insist that evaluation cannot properly take place without a complete specification of educational objective functions for every level of government, every type of program and target population, all regions, every type of student, and, indeed, for every educational decisionmaker.

50. This report has recommended a course of action different from all three. Although existing tests have shortcomings, some knowledge is better than none and therefore let us not abandon cognitive achievement measures. The mean is easy to use, but more knowledge is better than some, so we should go beyond simple averages. And although objective functions for evaluation are elegant, their practical application in education faces overwhelming obstacles.

51. The measures proposed here need further research before their exact properties are understood. *Which* exact statistics and which estimators to employ are open questions. As in the case of income distribution, there may be legitimate debate about which statistics are best. But also as in that case, the argument is that *some* such measures are better than none.

52. How should these statistics be used in the near term? Crude measures should be employed crudely. Continuous, cardinal uses of the

statistics proposed would probably mislead more than they would help. Percentile transformations of each statistic would be an improvement, since percentiles merely rank schools ordinally. But a further move away from pseudo-exactness is advisable. One might divide each percentile measure into five or so categories (say, the highest 20 percent of schools on each measure would receive a one, the second 20 percent a two, and so forth). (See also Dyer, 1972.) One might then envision a scheme like that shown in Table 4.

One should resist the temptation to concoct a grand measure, some weighted sum of all ten suggested statistics, and then to impose it on the evaluation process. Weighted sums assume mutual preferential independence, which does not hold for the distortion and proportion measures mathematically, and probably does not hold (given most reasonable objective functions) for any of the measures. To take a comparable example: How one feels about income distribution probably depends on the general level of a country's income (Rescher, 1966, pp. 36 ff); and any evaluation of the poverty line is in part a function of who the particular individuals are that fall beneath it. Similarly, assessing the intra-school spread is probably not advisable without considering the mean; and the existence of underachievers may bother one more if they are predominantly members of one ethnic or socioeconomic grouping. Although complicated algorithms expressing conditional preferences are possible, it is best not to include these formally in any data system, accountability scheme, or large-scale evaluation. Let each decisionmaker (and each citizen) be his own judge.

53. To propose the introduction of measures without clear-cut objectives flies in the face of rationalist predispositions. But new measures, even imperfect ones, can be the first step toward educational change. James March has suggested that most rethinking of objectives that does take place in organizations occurs precisely in a "backward" fashion--from changes in measures and performance indicators to changes in bureaucratic goals and operations.

Table 4

EXAMPLE OF THE USE OF NEW ACHIEVEMENT STATISTICS

Educational Objective	Achievement Measure	School Number				
		101	102	103	104	etc.
General achievement level	Mean	2	5	3	3	
Achievement relative to student background	Residual mean	4	3	1	4	
Equality of achievement	Spread (perhaps σ)	1	3	2	4	
Equalizing effect of school	Actual minus expected spread	3	1	2	2	
Mobility afforded by school	Residual spread	2	4	2	5	
Effectiveness with exceptional children	Distortion (perhaps skewness)	1	3	2	5	
Effectiveness with over- and underachievers	Residual distortion	3	5	1	3	
Assuring children achievement skills at minimum level K	Proportion of students ($A \geq K$)	2	4	2	4	
Assuring children do not underachieve below level C	Proportion of students ($R \geq C$)	5	1	4	1	
Success with children above (below) background level S	Mean score of students above (below) S	3	3	1	5	

Numbers under schools refer to the following table:

Percentile	Category
80-100	1
60-80	2
40-60	3
20-40	4
0-20	5

Percentiles are computed for each statistic.

The chief benefit from using the new statistics might be in changing institutions, rather than in producing more efficient evaluations (which should also occur). Measures inevitably become standards of performance. When measures are simplistic or partial, they can engender adverse incentives in the system. To beat the numbers game, any number of nonproductive gambits may evolve. And it is easier to create adverse incentives with just one measure (the mean) than with the ten new ones.

Using new statistics may shift discussions between educators and evaluators from questions of overall levels of performance to questions of equity, mobility, special programs, and the rest. This shift would not only more faithfully reflect the multiple and varied nature of educational objectives, it might also stimulate new concerns and create new incentives for action (or avoid some unwelcome old ones).

54. Further research may indicate other ways of obtaining more useful measures to serve the purposes of those proposed here. Stratifications of the data, for example, may produce desirable indicators of change scores across many categories of students and schools (Bruno, 1972); which method is preferable may hinge on the simplicity of each technique's application. Going beyond unadjusted achievement scores to residual measures places great strain on statistics of nonschool background. Residual measures are also sensitive to errors of misspecification, choice of fit, student turnover and attrition, and other problems discussed earlier. Nonetheless, the statistics in Table 4 are promising proxies for such basic educational goals as opportunity, mobility, equality, and differential preferences for different levels of achievement. Pending the badly needed research on educational objectives and their measurement, the introduction of these evaluative statistics, even with their considerable uncertainties and ambiguities, may enable policy makers to get more mileage from existing data.

REFERENCES

Acland, Henry D., "School Effects: An Analysis of New York Elementary Schools," 1972 (unpublished).

Andrews, D. F. *et al.*, *Robust Estimates of Location*, Princeton University Press, 1972.

Arditti, Fred D., "Risk and the Required Return on Equity," *Journal of Finance*, Vol. 22, No. 1, March 1967, pp. 19-36.

Arrow, Kenneth J., *Social Choice and Individual Values*, Wiley, New York, 1951.

Averch, Harvey A. *et al.*, *How Effective is Schooling?* R-956-PCSF/RC, The Rand Corporation, March 1972.

Barro, Stephen M., "An Approach to Developing Accountability Measures for the Public Schools," *Phi Delta Kappan*, Vol. 52, No. 4, December 1970, pp. 196-205.

Bauer, Raymond A., "Detection and Anticipation of Impact: The Nature of the Task," in R. A. Bauer, ed., *Social Indicators*, MIT Press, Cambridge, Mass., 1966.

Bickel, P. J., "On Some Analogues to Linear Combinations of Order Statistics in the Linear Model," *The Annals of Statistics*, Vol. 1, No. 4, July 1973, pp. 597-616.

Bowles, Samuel and Henry M. Levin, "The Determinants of Scholastic Achievement--An Appraisal of Some Recent Evidence," *Journal of Human Resources*, Vol. 3, No. 1, Winter 1968, pp. 2-24.

Brown, Byron W., "Achievement, Costs and the Demand for Public Education," *Western Economic Journal*, Vol. 10, No. 2, June 1972, pp. 198-219.

Brownlee, K. A., *Statistical Theory and Methodology in Science and Engineering*, Second Edition, Wiley, New York, 1965.

Bruno, James E., "A Methodology for the Evaluation of Instruction or Performance Contracts Which Incorporates School District Utilities and Goals," *American Educational Research Journal*, Vol. 9, No. 2, Spring 1972, pp. 175-195.

Chipman, John S., "The Ordering of Portfolios in Terms of Mean and Variance," *Review of Economic Studies*, Vol. 40 (2), No. 122, March 1973, pp. 167-190.

Craegar, John A., "Orthogonal and Nonorthogonal Methods for Partitioning Regression Variance," *American Educational Research Journal*, Vol. 8, No. 4, November 1971, pp. 671-676.

Cronbach, Lee J. and Lita Furby, "How We Should Measure 'Change'--Or Should We?" *Psychological Bulletin*, Vol. 74, No. 1, July 1970, pp. 68-90.

Dyer, Henry S., "The Measurement of Educational Opportunity," in Frederick Mosteller and Daniel P. Moynihan, eds., *On Equality of Educational Opportunity*, Random House (Vintage Books), New York, 1972, pp. 513-527.

Dyer, Henry S., Robert L. Linn, and Michael J. Patton, "A Comparison of Four Methods of Obtaining Discrepancy Measures Based on Observed and Predicted School System Means on Achievement Tests," *American Education Research Journal*, Vol. 6, No. 4, November 1969, pp. 591-605.

Fisher, Irving N. and George R. Hall, "Risk and Corporate Rates of Return," *Quarterly Journal of Economics*, Vol. 83, No. 1, February 1969, pp. 79-92.

Forsyth, Robert A., "Some Empirical Results Related to the Stability of Performance Indicators in Dyer's Student Change Model of an Educational System," *Journal of Educational Measurement*, Vol. 10, No. 1, Spring 1973, pp. 7-12.

Friedman, Milton and Leonard J. Savage, "The Utility Analysis of Choices Involving Risk," *Journal of Political Economy*, Vol. 56, No. 4, August 1948, pp. 279-304.

Guba, Egon G., "Development, Diffusion, and Evaluation," in Terry L. Eidell and Joanne M. Kitchel, eds., *Knowledge Production and Utilization in Educational Administration*, ERIC: ED 024 112, 1967.

Guthrie, James W., "What the Coleman Reanalysis Didn't Tell Us," *Saturday Review*, July 22, 1972.

Hanushek, Eric A. and John F. Kain, "On the Value of *Equality of Educational Opportunity* as a Guide to Public Policy," in Frederick Mosteller and Daniel P. Moynihan, eds., *On Equality of Educational Opportunity*, Vintage, New York, 1972, pp. 116-146.

Hicks, John R., *Value and Capital*, Second Edition, Oxford University Press, London, 1950.

Høivik, Tord, "Social Inequality--The Main Issues," *Journal of Peace Research*, 1971, Vol. 2, pp. 119-142.

Jencks, Christopher S. *et al.*, *Inequality*, Basic Books, New York, 1972.

Klitgaard, Robert E. and George R. Hall, *A Statistical Search for Unusually Effective Schools*, R-1210-CC/RC, The Rand Corporation, March 1973.

Kristol, Irving, "About Equality," *Commentary*, Vol. 54, No. 5, November 1972, pp. 41-47.

Lindquist, E. F. and A. N. Hieronymus, *Iowa Tests of Basic Skills: Manual for Administrators, Supervisors, and Counselors*, Houghton Mifflin, Boston, 1964.

Lydall, Harold F., *The Structure of Earnings*, Oxford University Press, Oxford, 1968.

Mayeske, George W. *et al.*, *A Study of Our Nation's Schools*, U.S. Department of Health, Education and Welfare, Office of Education, Washington, D.C., 1969.

McLaughlin, Milbrey Wallin, *Evaluation and Reform: The Case of ESEA Title I*, Ed.D. dissertation, Graduate School of Education, Harvard University, 1973.

Raiffa, Howard, *Decision Analysis*, Addison-Wesley, Reading, Mass., 1968.

Raiffa, Howard, *Preferences for Multi-Attributed Alternatives*, RM-5868-DOT/RC, The Rand Corporation, April 1969.

Raiffa, Howard, *Tradeoffs Under Certainty*, 1971 (unpublished).

Rescher, Nicholas, *Distributive Justice*, Bobbs-Merrill, Indianapolis, 1966.

Roche, J. G., *Investigation of Cost-Benefit and Decision-Analytic Techniques in Local Education Decisionmaking*, D.B.A. dissertation, Graduate School of Business Administration, Harvard University, 1971.

Stevens, William K., "Test Expert Calls I.Q. and Grade Equivalency Scores 'Monstrosities,'" New York *Times*, March 23, 1971.

Suchman, Edward A., *Evaluative Research*, Russell Sage, New York, 1967.

Tintner, Gerhard, "A Contribution to the Non-Static Theory of Choice," *Quarterly Journal of Economics*, Vol. 56, No. 2, February 1942, pp. 274-306.

Tukey, John W., *Exploratory Data Analysis* (limited preliminary edition, 3 vols.), Addison-Wesley, Reading, Mass., 1970.

Vaizey, John, *The Economics of Education*, Free Press, New York, 1962.

von Neumann, John and Oskar Morgenstern, *Theory of Games and Economic Behavior*, Princeton University Press, 1944.

Wargo, Michael J., Peggie L. Campeau, and G. Kasten Tallmadge, *Further Examination of Exemplary Programs for Educating Disadvantaged Children*, AIR-2026-7/71, American Institutes for Research, 1971.

Wilks, Samuel S., *Mathematical Statistics*, Wiley, New York, 1962.

Yarnold, James K., "The Minimum Expectation in χ^2 Goodness of Fit Tests and the Accuracy of Approximations for the Null Distribution," *Journal of the American Statistical Association*, Vol. 65, No. 330, June 1970, pp. 864-886.